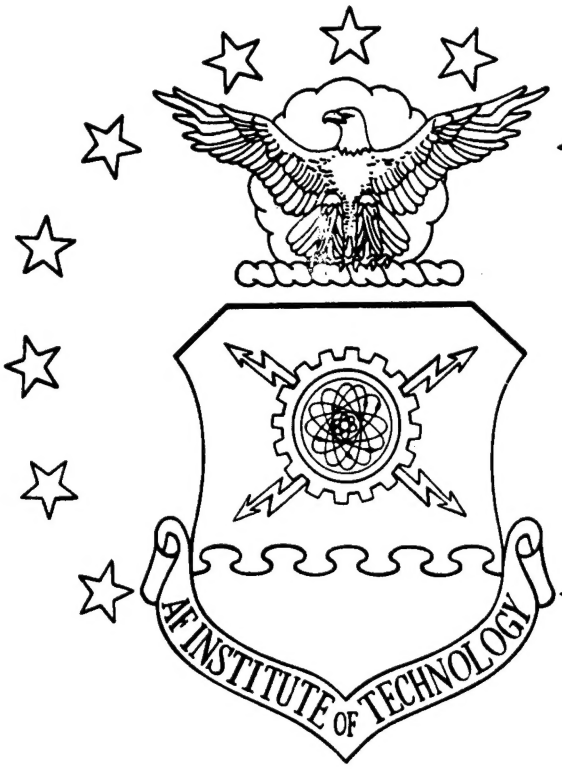


THE MATHEMATICS OF MEASURING CAPABILITIES OF ARTIFICIAL NEURAL NETWORKS

DISSERTATION

Martha Ayers Alvey Carter
Mathematician, USAF

AFIT/DS/ENC/95J-01

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

J64

AFIT/DS/ENC/95J-01

THE MATHEMATICS OF MEASURING CAPABILITIES OF ARTIFICIAL NEURAL NETWORKS

DISSERTATION

Martha Ayers Alvey Carter
Mathematician, USAF

AFIT/DS/ENC/95J-01

AFIT/DS/ENC/95J-01

THE MATHEMATICS OF MEASURING CAPABILITIES OF ARTIFICIAL NEURAL NETWORKS

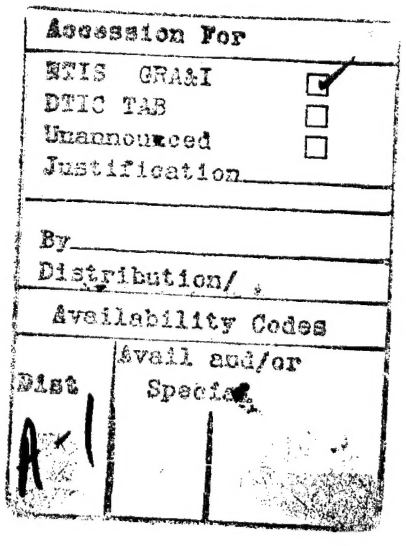
DISSERTATION

Presented to the Faculty of the Graduate School of Engineering
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Martha Ayers Alvey Carter, B.S., M.S.
Mathematician, USAF

June, 1995

Distribution unlimited



AFIT/DS/ENC/95J-01

THE MATHEMATICS OF MEASURING CAPABILITIES OF ARTIFICIAL NEURAL NETWORKS

Martha Ayers Alvey Carter, B.S., M.S.

Mathematician, USAF

Approved:

Mark E. Oxley 5 June 1995

Dr. Mark E. Oxley, Research Advisor

5 JUNE 1995

Dr. Gregory T. Warhola

5 June 1995

Dr. Steven K. Rogers

Dennis W. Ruck 30 MAY 1995

Dr. Dennis W. Ruck

5 Jan 95

Dr. William P. Baker, Dean's Representative

Robert A. Calico, Jr

Dean, Graduate School of Engineering

Preface

The ideas for this research were born out of the desire to be able to explain the mechanics of neural networks mathematically. I first learned of neural networks in 1989 in a Pattern Recognition class taught by Dr. Matt Kabrisky and Dr. Steve Rogers at AFIT. They delivered the material in a way that tapped my curiosity to discover what these things really were. I pursued this curiosity as a Ph.D. dissertation topic with their encouragement, for which I am especially grateful. Hopefully, it has resulted in useful research.

The task of guiding me through the research was taken on by Dr. Mark E. Oxley. I want to express my deep appreciation for his investment in my entire Ph.D. program. Dr. Oxley was a tremendous partner for this project. He willingly let me lead the research beyond the scope of our expertise. This resulted in monumental effort, on both our parts, to learn new mathematics and the intricacies of the field of artificial neural networks. Dr. Oxley was tenacious and offered constant enthusiasm. Moreover, he patiently taught me methods of research that I'm sure will prove useful in the future.

I would also like to thank the rest of my committee Maj Greg Warhola, Dr. Steve Rogers and Maj Dennis Ruck for serving as a great sounding board. Additionally, I want to thank Dr. William Baker for being a thorough and patient Dean's Representative. Special thanks go to the National Air Intelligence Center for supporting my desire to continue my education.

And finally, for his love, patience, understanding and encouragement, I am most grateful and forever indebted to my best friend and husband, Doug Carter. This could not have been done without him.

Martha Ayers Alvey Carter

Table of Contents

List of Figures

Abstract

Researchers rely on the mathematics of Vapnik and Chervonenkis to capture quantitatively the capabilities of specific artificial neural network (ANN) architectures. The quantifier is known as the V-C dimension, and is defined on functions or sets. Its value is the largest cardinality l of a set of vectors in $\mathbf{R}^d$ such that there is at least one set of vectors of cardinality l such that all dichotomies of that set into two sets can be implemented by the function or set. Stated another way, the V-C dimension of a set of functions is the largest cardinality of a set, such that there exists one set of that cardinality which can be shattered by the set of functions. A set of functions is said to shatter a set if each dichotomy of that set can be implemented by a function in the set. There is an abundance of research on determining the value of V-C dimensions of ANNs. In this document, research on V-C dimension is refined and extended yielding formulas for evaluating V-C dimension for the set of functions representable by a feed-forward, single hidden-layer perceptron artificial neural network.

The fundamental thesis of this research is that the V-C dimension is not an appropriate quantifier of ANN capabilities. Consequently, the results of this research provide a basis of mathematics on which to build quantifiers that address the specifics of ANN's ability based on invariant characterizations of signed sets. Specifically, the lattice structure of ANNs is investigated. Moreover, a cut-intersection semi-lattice is established upon which invariant analysis of an arrangement of hyperplanes can be examined. As a consequence of the study of combinatorial geometry of hyperplane arrangements, it is shown that solutions to the *chamber counting problem* that are based on analysis of the Poincaré polynomial also provide a closed form relation for determining the value of the V-C dimension of ANNs. This provides a relationship

between the study of combinatorial geometry of hyperplane arrangements and ANN capability analysis.

In addition, a generalized framework in which to perform ANN capabilities analysis is presented. The framework is based on invariant analysis. Moreover, an invariant based on geometric complexity defined by concepts of combinatorial geometry is presented and evaluated.

Finally, an instantiation of the framework is given. The quantifier is called the Ox-Cart dimension. It is a function of an invariant called the geometric complexity. This quantifier is directed at analysis of specific geometric arrangements of signed sets. In other words, O-C dimension characterizes an ANN's ability to solve a particular classification problem where the problem is characterized by its geometric complexity. This differs from V-C dimension which is about arbitrary sets. V-C dimension characterizes an ANN's ability to solve the worst case geometry and dichotomy of a classification problem characterized only be cardinality. Using the Ox-Cart dimension a lattice of feed-forward, single hidden-layer, perceptron artificial neural network is defined. This structure is shown to facilitate ANN architecture capabilities comparisons.

THE MATHEMATICS OF MEASURING CAPABILITIES OF ARTIFICIAL NEURAL NETWORKS

I. Introduction

The research described in this document is dedicated to improving the methods and mathematics used to determine capabilities of feed-forward artificial neural networks (ANNs). A generalized framework for deriving ANN capability quantifiers is developed. In order to provide the framework, the mathematical structure of the problem, generalized solution, and methods of analyzing both had to be clearly defined and investigated. *Invariant analysis* is used extensively to define the generalized framework. As a result, a capability quantifier called the *Ox-Cart dimension* is defined which characterizes the adequacies of ANN solutions to the classification problem. The Ox-Cart dimension is based on a geometric characterization of classification problems, called the *geometric complexity.* Both the Ox-Cart dimension and the geometric complexity exemplify structures in the generalized framework and exhibit desired invariant properties. The Ox-Cart dimension will be used to define a partial ordering which will produce the lattice of feed-forward, single hidden-layer, perceptron ANNs. The lattice structure facilitates comparisons of ANN architectures based on their ability to solve classification problems.

The novelty of this approach to quantify ANNs' capabilities is that it is designed to be consistent with properties specific for analyzing neural network architectures and problem specific characteristics. This is different from current methods based on the *Vapnik-Chervonenkis (V-C) dimension* which analyzes an ANN's ability to solve *arbitrary* classification problems. However, in some aspects the Ox-Cart dimension parallels V-C dimension. Hence, the repeated use of the term dimension.

(It should be noted that the term Ox-Cart dimension was created solely to parallel V-C dimension. Neither quantifiers are dimensions in the traditional mathematical sense, the cardinality of a basis set.)

Considerable background research on the V-C dimension approach led to significant results in two general topics. The first result extended existing bounds and precise values for V-C dimension and related mappings. This was accomplished by analyzing and exploiting the algebraic and topological structures of the mappings. The second result found that there is a strong connection between the study of ANNs' capability and an area of mathematics that will be referred to as combinatorial geometry. In fact, it will be shown that the V-C dimension of an ANN can be evaluated using combinatorial geometric methods for counting chambers of a hyperplane arrangement.

Combinatorial geometry is a rich area of mathematics that relies heavily on a lattice structure. When that structure is in place, much can be said about the capacity of a system; specific to this research are the systems generated by ANNs. In this document, the groundwork will be laid to make combinatorial geometry available to the artificial neural network community of researchers. Specifically, the *cut-intersection semi-lattice* of an ANN will be defined and used to obtain invariants for capability analysis. Additionally a lattice of the chamber sets of ANNs will be established.

This research relied heavily on previous work in the area of measuring ANN capabilities. For example, extending the results of the significant 1957 work of Andrei Kolmogorov, G. Cybenko showed that the set of functions that can be approximated by a multi-layer perceptron network are dense in the set of continuous functions defined on the n-dimensional cube (18). This result was extended further by A. Ronald Gallant and Halbert White who showed that ANNs are dense in Sobolev Spaces, suggesting that derivative information can be approximated (19).

What is not so clear about the capabilities of an ANN is what it takes to learn these functions. What size ANN is required? How many training samples are needed? Recent publications on this subject have borrowed from important works by Cover and Vapnik-Chervonenkis. Baum used theorems of Cover to give bounds on the required ANN size for certain problems. More recently, Eduardo Sontag proposed mappings similar to V-C dimension for the specific purpose of analyzing ANN capabilities. Sontag derived interesting results about sigmoidal transfer functions and direct connections using his mappings (39). This dissertation extends these ideas even further.

The following sections will explain the details of classification problems, artificial neural networks, and what is meant by the ability to generalize. With this explanation, a common set of terminology will be established.

1.1 Problem: Generalization of Classification Data

In general, there are two distinct problems solved with artificial neural networks: function approximation (interpolation) and data classification. One can think of both of these problems as function approximation, where interpolation approximates functions of the form

$$f : \mathbf{R}^{d_1} \rightarrow \mathbf{R}^{d_2} \qquad d_1,\ d_2 \in \mathbf{N}$$

and two-class classification problems approximate functions of the form

$$f : \mathbf{R}^{d} \rightarrow \{0, 1\} \qquad d \in \mathbf{N}.$$

The distinction is made since analysis of complexity of the problems is different. (Note that throughout this document $\mathbf{R}^d$ refers to d-dimensional Euclidean space.) The research outlined by this document deals with the problem of *classification* of data. The research will concentrate on the *two-class problem.*

The two-class problem (sometimes called the two-coloring problem which is a special case of the n-coloring problem) refers to the problem of separating a set of points, $X \subset \mathbf{R}^d$ into two pre-defined subsets, X^+, X^- where $X^+ \cap X^- = \emptyset$, and $X^+ \cup X^- = X$. The two subsets, X^+ and X^-, together define a *dichotomy* or *partition* on X. It is said that the solution to a classification problem is an function f which *implements* the dichotomy, that is $f(x) > 0$ for all $x \in X^+$ and $f(x) < 0$ for all $x \in X^-$ for a given set X. The ordered pair (X^+, X^-) is referred to as a *signed set.* Referring to (X^+, X^-) as an ordered pair is meaningful and accurate since $(X^+, X^-) \neq (X^-, X^+)$. The *space of signed sets* is investigated in detail as part of the new research presented in this dissertation.

1.2 Tools: Feed-Forward, Single Hidden-Layer, Perceptron Artificial Neural Networks

The research in this dissertation concerns *feed-forward, single hidden-layer, perceptron artificial neural networks.* See Figure 1 for a graphical depiction. *Feed-forward* refers to the upward direction of the graphical depiction in which the input vectors are processed. Note that the picture shows three layers of processing units referred to as *nodes, units,* or *processors.* The bottom layer is the *input layer* of nodes. This is the starting point for the input data. The middle layer, is referred to as the *hidden-layer.* The top layer is referred to as the *output layer.* Each layer is connected by *interconnections.* Each interconnection has an associated *weight.* Each node performs a process on incoming vectors or scalars and outputs vectors or scalars. The process may differ depending on the layer. In a single hidden-layer perceptron artificial neural network, the middle layer nodes are *perceptron* nodes. In the case of a perceptron node, the input is usually the dot product of the input vector and the node's incoming weight vector. This dot product is compared to the node's threshold value, and based on the comparison a scalar is output. The *comparison* is performed by a *sigmoid function* sometimes referred to as a threshold

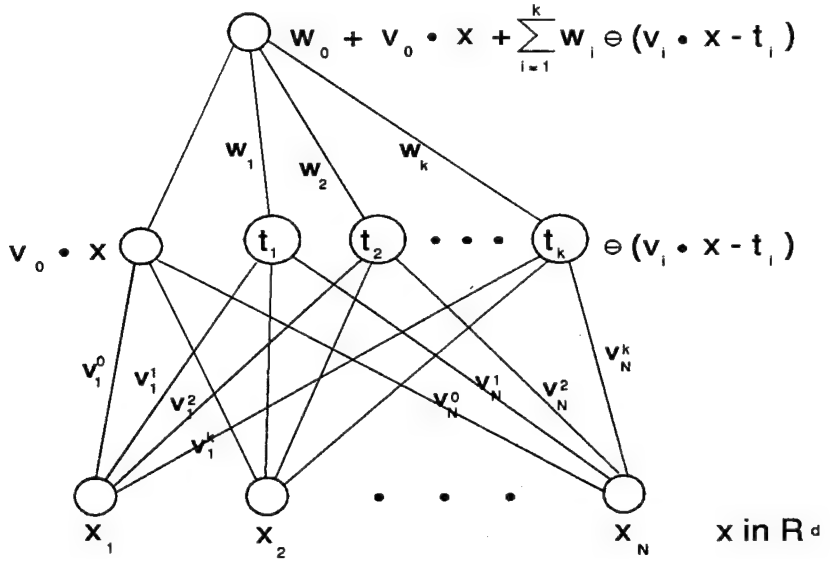


Figure 1. A feed-forward, single hidden-layer, perceptron ANN.

function or transfer function. In addition, a *direct connection* may be included. The direct connection processes the input vector straight to the output layer bypassing the middle layer and has no associated sigmoid function.

Now consider the mathematics of the ANN.

Definition. A function, $\theta : \mathbf{R} \to \mathbf{R}$, will be called a sigmoid if

$$t_+ := \lim_{s \to +\infty} \theta(s) \text{ and } t_- := \lim_{s \to -\infty} \theta(s).$$

The following are two examples of commonly used sigmoid functions. The first is the standard sigmoid function

$$\theta(s) = \frac{1}{1 + e^{-s}}. \tag{1}$$

The second is referred to as the hard-limiter or Heaviside function

$$H(s) = \begin{cases} 0 & \text{if} \quad s \leq 0 \\ 1 & \text{if} \quad s > 0. \end{cases} \tag{2}$$

The input into each perceptron node is the dot product of the input vector, $x \in \mathbf{R}^d$, and the node's associated incoming interconnection weight vector, $v_i \in \mathbf{R}^d$. Each node has an associated scalar threshold value, $\tau_i \in \mathbf{R}$. The threshold value is subtracted from the dot product and the result is passed through a sigmoid. The output of each perceptron can be defined as $\theta(v_i \cdot x - \tau_i)$. Note that if $f(x) = v_i \cdot x - \tau_i$, then $f(x) = 0$ is the equation of a *hyperplane* in $\mathbf{R}^d$. The function f is also referred to as a *separating surface.*

Each of the k perceptron output values, which are scalar, are multiplied by the perceptron's associated outgoing interconnection weight, ω_i, which is a scalar. These values are the input to the output layer where they are summed together to produce

$$\sum_{i=1}^{k} \omega_i \theta(v_i \cdot x - \tau_i).$$

Define the sigmoid function $\mathcal{H}$ to be

$$\mathcal{H}(s) = H(s) - H(-s) = \begin{cases} 1 & \text{if} \quad s > 0 \\ 0 & \text{if} \quad s = 0 \\ -1 & \text{if} \quad s < 0. \end{cases}$$

Note that if $\theta = \mathcal{H}$, then the value $\mathcal{H}(v_i \cdot x - \tau_i)$ can be interpreted as an indication of the input vector, x, being on one side of the hyperplane, $\{x \in \mathbf{R}^d : f(x) = v_i \cdot x - \tau_i = 0\}$, or the other. Equivalently, it indicates if x is in the *positive halfspace,* h^+, where

$$h^+ = \{x : f(x) > 0\}$$

or the *negative halfspace*, h^-, where

$$h^- = \{x : f(x) < 0\}.$$

Moreover, if the outgoing weights are binary, i.e. $\omega_i \in \{0, 1\}$, then the value

$$\sum_{i=1}^{k} \omega_i \theta(v_i \cdot x - \tau_i)$$

can distinguish between *intersections* of the halfspaces. Moreover, it completely defines the function F of an ANN for any $x \in \mathbf{R}^d$. Specifically, define F as

$$F(x) = \sum_{i=1}^{k} \omega_i \theta(v_i \cdot x - \tau_i), \tag{3}$$

where $\omega_i \in \{0, 1\}$, $\theta = \mathcal{H}$, $v_i \in \mathbf{R}^d$, and $\tau_i \in \mathbf{R}$ for $i = 1, \ldots, k$. This is the equation of the ANN which is the object of capability analysis of the new research in this dissertation.

Equation 3 is not the most general definition of a feed-forward, single hidden-layer, perceptron artificial neural network. The most general case requires the perceptron's outgoing interconnection weights to be any real value. Additionally, a *direct connection* is included that is defined by

$$\omega_0 + v_0 \cdot x,$$

where $\omega_0 \in \mathbf{R}$ and $v_0 \in \mathbf{R}^d$. Combining yields the following function, G, that defines a general feed-forward, single hidden-layer, perceptron artificial neural network. Specifically,

$$G(x) = \omega_0 + v_0 \cdot x + \sum_{i=1}^{k} \omega_i \theta(v_i \cdot x - \tau_i), \tag{4}$$

where θ is any sigmoid function, $v_i \in \mathbf{R}^d$, $\omega_i, \tau_i \in \mathbf{R}$, for $i = 0, 1, \ldots, k$. Note that $F = G$ if $\omega_i \in \{0, 1\}$, $\theta = \mathcal{H}$, $\omega_0 = 0$, and $v_0 = 0$.

The parameters v_i, ω_i, and τ_i, for $i = 0, 1, \ldots, k$ are approximated to yield an approximate function. The approximated function will be referred to as a *solution* or an *ANN solution* to the classification problem described by a signed set. The approximation process is referred to as *learning* or *training* and is accomplished with the use of a *training set* of data which is a signed set. All possible such G's will be referred to as a family.

1.3 Analysis: Generalization

In this research, there are basically two questions being asked about an artificial neural network's ability to *generalize* about classification data. One is: *What size net is required to accomplish a given classification problem?* That is, *What is the value of k?* The second question is: *How large of a set of vectors can a given architecture dichotomize?*. This division of perspectives is evidenced in the literature described in Chapter 2. Both of these questions and their relationship will be partially answered in this research.

It is important to clarify the term *generalization.* In this document, for an ANN to generalize a signed set, (X^+, X^-), means that after training the ANN has produced a function G that implements the dichotomies of the training set (which is a signed set). Hence, the analysis of the capabilities of a family of functions, G, would be an investigation of an arbitrary G's ability to implement all dichotomies in the training set. This is also referred to as the family's *separating capacity.* It should be noted that the term generalization is sometimes used differently in the literature. The definition presented above could be referred to as *memorization,* in which case the analysis of the performance of G to generalize would be an investigation of G's performance on data not used for training.

1.4 Synthesis: Achieving Generalization

Achieving generalization, i.e., approximating the parameters v_i, ω_i, and τ_i, for $i = 0, 1, \ldots, k$ is performed by the chosen *learning algorithm.* The performance of learning algorithms is not addressed by this research. Rather, there is an assumption that if a solution (best choice of parameters) exists, then it can be learned. This assumption is supported by Kolmogorov, Hecht-Nielsen and Cybenko (24) (18).

1.5 Validation: Testing the Results

Validation is the portion of the problem solving process that deals with testing the results and determining its validity. Validation is not addressed directly since the scope of this research does not involve a particular application and its solution. This should not be confused with validating the adequacy of the proposed ANN capability quantifiers which will be addressed.

1.6 Conclusions

In summary, this chapter provided a basis of general knowledge about solving classification (two-class) problems using ANNs. It also explained what portions of the solution process are directly addressed as the new research in this dissertation.

The second chapter details relevant literature to motivate technically the new research presented in the following chapters. Chapter III provides a re-evaluation of existing results which motivated extensions of that work. The fourth chapter discusses desired properties of capability quantifiers lacking in V-C based quantifiers. Chapter V will introduce basics of lattice theory and combinatorial geometry, and recast the mathematics of ANNs in the context of lattice theory. Chapter VI will present a generalized framework in which ANN capability analysis can be performed. Chapter VII will bring together the innovations of the generalized framework of Chapter VI and the rich mathematics of Chapter V to define the Ox-Cart dimension

and the lattice of ANNs. Ideas for follow-on research will be enumerated in Chapter VIII. A summary is provided in Chapter IX.

II. Background

This chapter reviews research on the problem of determining capabilities of certain artificial neural networks (ANNs). In particular, the architectural requirements for an ANN to have the capability to approximate arbitrary functions is addressed. Included are methods to quantify the capability of a class of separating sets and explanations on how these quantifiers direct the approach to solving classification problems. The material presented in this chapter is included to motivate the research presented in this dissertation. In particular, the material outlines a technical progression of research which serves as a catalyst for the results given in the following chapters.

This chapter discusses published research on the Vapnik-Chervonenkis (V-C) dimension of a set of separating surfaces. This quantifier or mapping (sometimes inaccurately referred to as a measure in the literature), also known as the separating capacity of a class of sets, directs the determination of the size of the ANN required to solve a problem (9). Valiant's work emphasizes the importance of an accurate ANN capability quantifier by showing that the V-C dimension also directs the amount of training data required for adequate learning (45). Learning is not discussed in detail in this document. Hence, Valiant's results are not presented in detail. However, Valiant's results are mentioned as testimony for the usefulness of research related to refining the definition of V-C dimension and providing bounds or values for the quantifier, given particular ANN architectures.

The technical review provided here, includes Cover's work about a single separating surface which could be a hyperplane defined by a perceptron or a nonlinear surface. Also included are Baum's results about multiple separating hyperplanes defined by multiple *perceptrons* in a feed-forward single hidden-layer perceptron ANN. Additionally, Sontag's results about alternative V-C based quantifiers for ANNs are

presented. Before the results are discussed, two concise definitions of V-C dimension and how V-C dimension relates to ANNs are provided.

2.1 Vapnik-Chervonenkis Dimension

Vapnik-Chervonenkis (V-C) dimension is a mapping that describes the separating capacity of a set of sets. Put another way, V-C dimension is a quantifier of how well a set of sets can implement arbitrary dichotomies of a signed set. The set of sets which are the domain of the V-C dimension can simply be an arbitrary set of sets, or be derived from a set of functions. Since the set of sets can be derived from functions, V-C dimension is a useful quantifier of classification capability of ANNs. In particular, the set of sets generated by ANNs are the halfspaces and their intersections generated from the separating surfaces, the hyperplanes defined by perceptrons.

The following are two explicit definitions of V-C dimension. The first is a general definition defined on a set of sets and is easy to work with mathematically. The second is more specific since it is defined on a set of sets generated from separating surfaces.

Consider the following general definition of V-C dimension outlined by Wenocur and Dudley in (47). This definition is not as intuitive as the second, but will provide the required flexibility for the new research in this dissertation. Given a nonempty set, $X \subset \mathbf{R}^d$, a collection, Ω, of subsets of X, and a finite set $Y \subset X$, let $\Delta^{\Omega}(Y)$ denote the number of distinct sets $O \cap Y$ for $O \in \Omega$. That is

$$\Delta^{\Omega}(Y) = \text{card}(\{O \cap Y : O \in \Omega\}) \leq 2^n.$$

Define

$$m^{\Omega}(n) \equiv \max\{\Delta^{\Omega}(Y) : Y \subset X,\ \text{card}(Y) = n\} \leq 2^n.$$

Notice that $m^{\Omega}(n)$ is no larger than all possible dichotomies of Y, that is $m^{\Omega}(n) \leq 2^n$ for all n. Define $VC(\Omega)$ as the V-C dimension of the set of sets Ω given by

$$VC(\Omega) \equiv \inf\{n : m^{\Omega}(n) < 2^n\}.$$

with

$$VC(\Omega) \equiv +\infty \text{ if } m^{\Omega}(n) = 2^n \text{ for all } n.$$

Now, we can make a second definition which is an instantiation of the first. Consider the following preliminary definitions.

Definition. A *dichotomy* of a set, $X \subset \mathbf{R}^d$, is the partition of its elements into two disjoint subsets X^+ and X^- such that $X^+ \cup X^- = X$ and $X^+ \cap X^- = \emptyset$ and is denoted by the ordered pair (X^+, X^-). (X^+, X^-) is referred to as a *signed set.* If there are n elements in X, then there are 2^n possible dichotomies of X.

Definition. A *separating surface*, f, on a set $X \subset \mathbf{R}^d$ is a function that maps X to $\mathbf{R}$.

Definition. A dichotomy, (X^+, X^-), of a set $X \subset \mathbf{R}^d$ is *implemented* by a set, $\mathcal{F}$, of separating surfaces if there exists $f \in \mathcal{F}$ such that $f(x) > 0$ for all $x \in X^+$ and $f(x) < 0$ for all $x \in X^-$.

Definition. A set of separating surfaces, $\mathcal{F}$, *shatters* the set X if every possible dichotomy of X can be implemented by a separating surface $f \in \mathcal{F}$.

Definition. The set of vectors $X \subset \mathbf{R}^d$ are said to be in *general position* if every subset $Y \subset X$ of d or fewer vectors is a linearly independent set.

With the above definitions, V-C dimension can be defined as follows.

Definition. The *Vapnik-Chervonenkis (V-C) Dimension of a set of surfaces*, $\mathcal{F}$, is the largest integer n, such that there is at least one set in general position, $X \subset \mathbf{R}^d$ of cardinality n that can be shattered by $\mathcal{F}$.

In other words, V-C dimension of a set of functions representable by ANNs is the largest set of data which can be guaranteed to be implemented regardless of the classification of the points in the data set.

It is meaningful here to reiterate that when the set of surfaces, $\mathcal{F}$, referred to in the above definitions, is the set of hyperplanes defined by the perceptron nodes in an ANN's hidden layer, then the V-C dimension of $\mathcal{F}$, is a quantifier about the capability of that architecture to implement arbitrary dichotomies of a set of vectors. The value of the V-C dimension indicates the cardinality of a set of training vectors which can be guaranteed to be classified correctly by an ANN regardless of the dichotomy.

As an example in $\mathbf{R}^2$, Choose Ω to be the set of halfspaces generated from the hyperplane associated with a single perceptron node. The V-C dimension of the set, $\mathcal{F}$, comprised of the single line in two dimensions, or equivalently the set, Ω, comprised of the two halfspaces generated by the line, is three. This is because there is a set of three points in general position which can be implemented and there is no set of four points in general position which can be shattered. Consequently, it can be said that the V-C dimension of an ANN with a single perceptron with a hard-limiter sigmoid, which takes as its input vectors in $\mathbf{R}^2$ is three.

2.2 Cover: The Separating Capacity of a Surface

In this section, it will be established that the *separating capacity,* the number of vectors that can be separated, of a certain set of nonlinear separating surfaces having d degrees of freedom is $2d$ vectors (16). The separating capacity is another term used to express the ability of separating surfaces to dichotomize signed sets. Hence, Cover (16) equates separating capacity with the V-C dimension. That is, the V-C dimension of a certain set of nonlinear separating surfaces with d parameters is $2d$. The surfaces are derived from homogeneous functions, $f_w : \mathbf{R}^d \rightarrow \{-1, 0, 1\}$. The surface is defined as $\{x : f_w(x) = 0\}$, where w is an arbitrary vector of parameters

called the weight vector. The separating capacity of a set of surfaces as defined by Cover is the number of vectors in a set whose dichotomies can be implemented with a probability of 1/2. Probability concepts are introduced because the number of random dichotomies of n points in d dimensions which can be implemented with certainty is shown to have a cumulative binomial distribution.(16)

The vectors of X are assumed to be randomly distributed in a finite dimensional space. The vectors are assumed to be in general position. A dichotomy of X, (X^+, X^-), is said to be separable *relative to a set of surfaces*, $\mathcal{F}$, if there exists a surface, $f \in \mathcal{F}$, that separates the points in X^+ from those in X^-; i.e. there exists $f_w \in \mathcal{F}$ such that

$$f_w(x) = \begin{cases} 1 & \text{if} \quad w \cdot x > 0 \quad \forall\, x \in X^+ \\ 0 & \text{if} \quad w \cdot x = 0 \\ -1 & \text{if} \quad w \cdot x < 0 \quad \forall\, x \in X^-. \end{cases} \tag{5}$$

2.2.1 The Function Counting Theorem. The Function Counting Theorem answers the question: *How many homogeneously, linearly separable dichotomies of n points in d-dimensional space are there?* (A set is said to be *homogeneously,* linearly separable if there exists an f_w that satisfies Equation 5 and $f_w(0) = 0$.) Consider the following well-established theorem given here without proof.

Theorem 1 (The Function Counting Theorem) *(16:326) There are $C(n, d)$ homogeneously, linearly separable dichotomies of n points in general position in d-dimensional space where*

$$C(n, d) = 2 \sum_{k=0}^{d-1} \binom{n-1}{k}.$$

2.2.2 Generalizing to Arbitrary Surfaces. Now, consider similar arguments for arbitrary separating surfaces. Let $\mathcal{F}$ be a set of arbitrary separating surfaces. Let (X^+, X^-) be a signed set of n points. Let ϕ be a vector-valued mapping

$$\phi : \mathbf{R}^d \rightarrow \mathbf{R}^d,$$

where $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_d(x))$ and $x \in \mathbf{R}^d$. Assume that ϕ is dimension preserving.

Definition. The set $\{x : w \cdot \phi(x) = 0\}$ is referred to as a *ϕ-surface.*

Definition. A dichotomy of X, (X^+, X^-) is *ϕ-separable*, if there exists $w \in \mathbf{R}^d$ such that

$$\begin{aligned} w \cdot \phi(x) &> 0 \quad \forall\, x \in X^+ \\ w \cdot \phi(x) &< 0 \quad \forall\, x \in X^-. \end{aligned}$$

Definition. If ϕ is as defined above and for each $X = \{x_1, x_2, ..., x_n\} \subset \mathbf{R}^d$, then X is said to be in *ϕ-general position* if every k-element subset of the set $\phi[X] = \{\phi(x_1), \phi(x_2), \ldots, \phi(x_n)\}$, is linearly independent for all $k \leq d$ and $X \subset \mathbf{R}^d$.

The following lemma enables the Function Counting Theorem to extend to arbitrary surfaces, and also to surfaces constrained to pass through a set of points.

Lemma 1 *Let (X^+, X^-) be a signed set of points in $\mathbf{R}^d$ and let $y \notin X$ be a point other than the origin in $\mathbf{R}^d$. Then, the dichotomies $(X^+ \cup \{y\}, X^-)$ and $(X^+, X^- \cup \{y\})$ are both homogeneously linearly separable if and only if (X^+, X^-) are homogeneously linearly separable by a $(d-1)$-dimensional subspace containing y. (16:327)*

Geometrically speaking, Lemma 1 provides that a new point can be adjoined to either half of a linearly separable dichotomy to form two new separable dichotomies if and only if there exists a separating hyperplane containing the new point that separates the original dichotomy.

Sufficient background has been established to present Cover's major result.

Theorem 2 *If a ϕ-surface, $\{x \in X : w \cdot \phi(x) = 0\}$, is constrained to contain the set of points $Y = \{y_1, y_2, \ldots, y_k\}$, where $\{\phi(y_1), \phi(y_2), \ldots, \phi(y_k)\}$ is linearly independent*

and where the projection of $\phi(x_1)$, $\phi(x_2)$, ..., $\phi(x_n)$ onto the orthogonal subspace to the space spanned by $\{\phi(y_1), \phi(y_2), \ldots, \phi(y_k)\}$ is in general position, then there are $C(n, d-k)$ ϕ-separable dichotomies of X. (16:327).

2.2.3 Separability of Random Patterns. There are two different notions of randomness associated with the classification problem.

- The set X is fixed in position, but the vectors in X are classified independently with equal probability into one of two classes.
- The set X, itself, is randomly distributed in space and the desired dichotomization may be random or fixed.

Either way, the separability of the signed set becomes a random event independent of the dichotomy and the geometric configuration. This leads to two questions: *What is the probability of being able to implement an arbitrary dichotomy?* and *What is the maximum number of points which can be separated by a family of ϕ-surfaces?*

Suppose that $X = \{x_1, x_2, \ldots, x_n\}$ is fixed and a dichotomy is chosen at random with equal probability from the 2^n possible dichotomies of X. Let X be in ϕ-general position with probability 1, and let $P(n, d)$ denote the probability that a random dichotomy is ϕ-separable. Again, d denotes the degrees of freedom of ϕ or equivalently the dimension of the image space under ϕ. Then, from Theorem 2, there are $C(n, d)$ ϕ-separable dichotomies of the 2^n total number of dichotomies. Hence,

$$P(n, d) = \frac{1}{2^n} C(n, d) = \frac{1}{2^{n-1}} \sum_{k=0}^{d-1} \binom{n-1}{k}. \tag{6}$$

Equation (6) is the cumulative binomial distribution corresponding to $d-1$ or fewer successes of $n-1$ trials where the event space is binary. This answers the first question.

The utility of Cover's material actually lies in the second question which is answered by Cover's following results. *What is the largest n such that $P(n, d) = 1$?*

(This is, in fact, the question also answered by the works of Baum and Sontag which are explained in later sections of this chapter.) In other words, what is the largest cardinality of a set such that any randomly selected dichotomy can be implemented with probability 1 by a ϕ-surface with d degrees of freedom.

Let $\{x_1, x_2, \ldots\}$ be a set of random vectors in general position and define the random variable, N, to be the largest integer such that $\{x_1, x_2, \ldots, x_N\}$ is ϕ-separable where ϕ has d degrees of freedom. Then, since Equation (6) is the cumulative binomial distribution, the probability that $N = n$ is the difference in the probabilities that $N \geq n$ and $N \geq n + 1$, or

$$\begin{aligned} P_r\{N = n\} &= P(n, d) - P(n + 1, d) \\ &= \left(\tfrac{1}{2}\right)^n \binom{n-1}{d-1}, \qquad n = 1, 2, \ldots \end{aligned}$$

This is commonly known as the negative binomial distribution with probability of failure equal to $\frac{1}{2}$. In this scenario, n is the number of trials required before d *failures* of a binary experiment are expected to be generated, and

$$\begin{aligned} E(n) &= 2d \\ Median(n) &= 2d. \end{aligned}$$

Here $E(\cdot)$ is the expected value operator. If d is chosen carefully with respect to n, then the asymptotic probability that n vectors are separable by a ϕ-surface of d degrees of freedom appears like the cumulative normal distribution. Specifically, let α be any real number then choose d approximately as follows: $d \approx \left(\frac{n}{2}\right) + \left(\frac{\alpha}{2}\right)\sqrt{n}$. Then,

$$P\left(n, \frac{n}{2} + \frac{\alpha}{2}\sqrt{n}\right) \sim \Phi(\alpha),$$

as $n \to \infty$, where $\Phi(\alpha)$ is the cumulative normal distribution

$$\Phi(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\xi^2/2} d\xi.$$

In other words, $\frac{P}{\Phi}$ approaches 1 asymptotically as n grows large.

The most interesting observation about $P(\cdot,\cdot)$ is that for $0 < \epsilon < 1$ and for all $d \in \mathbf{N}$,

$$\begin{aligned} \lim_{d\to\infty} P(2d(1+\epsilon), d) &= 0 \\ P(2d, d) &= \tfrac{1}{2} \\ \lim_{d\to\infty} P(2d(1-\epsilon), d) &= 1. \end{aligned}$$

This was shown by Winder (49). The threshold effect, where the number of vectors equals twice the number of degrees of freedom of the separating surface, suggests that $2d$ is the *separating capacity* of a set of separating surfaces having d degrees of freedom (16:331). Bringing this section together is the fact that this *separating capacity*, the maximum cardinality of a set whose random dichotomies can be implemented with probability 1 by a set of separating surfaces having d degrees of freedom, is the *V-C dimension* of the set of separating surfaces.

An important caveat to remember about Cover's work is that it is about the abilities of *one* separating surface from a particular family to separate a dichotomy. The next section will outline research that extended Cover's work to multiple separating surfaces for the purpose of quantifying artificial neural network capabilities.

2.3 Baum: Artificial Neural Network Architecture Size for Arbitrary Classification

In the last section, the separating ability of one surface was established. In this section, the separating ability of multiple surfaces will be investigated. In particular, the results of Baum (9) show that the V-C dimension of a feed-forward, single hidden-layer, perceptron ANN is finite. Moreover,

$$2\left\lfloor \frac{N_l}{2} \right\rfloor d \leq VC \leq 2N_w \log_2(eN_N),$$

where

N_l = number of hidden-layer nodes

d = dimension of input space

N_w = total number of weights in the ANN

N_N = total number of nodes in the ANN

e = Euler's number

VC = V-C dimension of the ANN.

Note that the lower bound is approximately equal to the number of weights connecting the input layer to the hidden layer of the ANN. The upper bound is not much more than twice the number of weights in the ANN. Hence, a rough estimate of the V-C dimension of the ANN could be the total number of weights in the ANN. Vapnik and Chervonenkis' results suggest that the number of vectors in the training set required for learning is at least the value of the V-C dimension (46). The background of this result is what is pertinent to the research proposed in this document. The extension of Cover's work to feed-forward, single hidden-layer, perceptron ANNs provides an interesting perspective on the capabilities of ANNs (9).

In the following lemma and theorem provided by Baum, the ANN is a feed-forward, single hidden-layer, perceptron ANN where the perceptrons have hard-limiters as the sigmoid function. The set to be shattered, X, has n points which are in general position in Euclidean d-space. The lemma establishes, by counter-example, that there must be n/d perceptron units in order to shatter the n points. The theorem provides the sufficiency condition for an ANN to be able to shatter a set of n points.

Lemma 2 *Any net capable of arbitrary dichotomies of n points in d dimensions must have at least n/d nodes in its hidden-layer. (9:198)*

The next theorem proves that a single hidden-layer ANN can in fact achieve arbitrary dichotomies with $\left\lceil \frac{n}{d} \right\rceil$ hidden-layer nodes.

Theorem 3 *A feed-forward, single hidden-layer, perceptron ANN with* $\left\lceil \frac{n}{d} \right\rceil$ *hidden-layer units can compute an arbitrary dichotomy on n d-dimensional vectors in general position. (9:199)*

Through counting arguments, Baum also establishes lower bounds on the number of hidden-layer nodes and the number of connection weights. In other words, the lower bounds are bounds under which there is no guarantee that an arbitrary dichotomy can be separated. Additionally, Baum consistently uses the notion of a worst-case geometric arrangement of a signed set to prove his results. The worst-case arrangement can be described as an n-gon in $\mathbf{R}^2$, where the vertices are the points in the signed set and have an alternating sign. (9:200-204)

In summary, Baum has specified the domain of V-C dimension to a set of functions computable by artificial neural networks, specifically feed-forward, single hidden-layer, perceptron ANNs with hard-limiters at each node. Moreover, he uses this to answer the question: *How many perceptron nodes are required to guarantee shattering the sets and how many training samples are required for adequate learning?*

2.4 Sontag: Alternative Classification Capabilities Quantifiers

The research that is described in this section addresses two questions that were begging to be answered in Baum's work. One is: *What if a standard sigmoid function is used at each node instead of the Heaviside function?* The other question is: *Is V-C dimension the appropriate measurement of ANN capability and if not, what is?* In fact, Sontag shows that using a continuous sigmoid instead of a hard-limiter appears to double the neural network capabilities. A side result of Sontag's is that including a direct connection also improves the capability of an ANN if it is solving a classification problem but not if it is solving an interpolation problem. The theme here is to evaluate some of the nuances of certain artificial neural network architectures in terms of V-C dimension. However, there is more. Sontag raises an important point that is the basis of the new research described in this dissertation.

Sontag's point is that V-C dimension is not always the appropriate quantifier of ANN capabilities. (All of the major results in this section are drawn from Sontag (39).)

2.4.1 Sontag's Notation and Definitions. Recall previous definitions of dichotomy of a set X, shattering, and V-C dimension of a set of functions $\mathcal{F}$ (see section 2.1). Consider the following definitions of other mappings of capability which are based on V-C dimension.

Definition. Let $\mathcal{F}$ be a set of scalar-valued functions defined on $\mathbf{R}^d$. Define the mapping $\overline{\mu}(\mathcal{F})$ to be the largest integer $n \geq 1$ (possibly ∞) such that there is *at least one* set X of cardinality n in $\mathbf{R}^d$ which can be shattered by $\mathcal{F}$.

Definition. Let $\mathcal{F}$ be a set of scalar-valued functions defined on $\mathbf{R}^d$. Define the mapping $\underline{\mu}(\mathcal{F})$ to be the largest integer $n \geq 1$ (possibly ∞) such that *every* set X of cardinality n can be shattered.

The utility of both of these quantifiers should be evident. Note that the first mapping is the V-C dimension less one, that is $\overline{\mu}(\mathcal{F}) = VC(\mathcal{F}) - 1$ for any $\mathcal{F}$. The second mapping appears like a more appropriate quantifier for determining ANN capability since it is the cardinality of sets in which *all* sets are guaranteed to be "shatterable", not just *one* set. Both of these mappings are fairly extreme, however. Hence, Sontag suggests another quantifier which is a more robust version of $\underline{\mu}$.

Definition. Let $\mathcal{F}$ be a set of scalar-valued functions defined on $\mathbf{R}^d$. Define the mapping, $\mu(\mathcal{F})$, to be the largest integer $n \geq 1$ (possibly ∞) for which the set of sets that can be shattered by $\mathcal{F}$ is dense, in the sense that given every n-element set $X = \{x_1, x_2, \ldots, x_n\}$, there are points $\widetilde{x_i}$ arbitrarily close to their respective x_i such that $\widetilde{X} = \{\widetilde{x_1}, \widetilde{x_2}, \ldots, \widetilde{x_l}\}$ can be shattered by $\mathcal{F}$.

Note that this is a topological method of assuring data points are in general position. Also, note that $\underline{\mu}(\mathcal{F}) \leq \mu(\mathcal{F}) \leq \overline{\mu}(\mathcal{F})$.

As an example, let $\mathcal{F}$ be the set of affine functions on $\mathbf{R}^2$: $f(x) = ax_1 + bx_2 + c$, i.e.,

$$\mathcal{F} = \{f : \mathbf{R}^2 \to \mathbf{R}^1 \mid f(x) = ax_1 + bx_2 + c \ , \ a, b, c \in \mathbf{R}\}.$$

Any dichotomy of a set of three points, which are not collinear in $\mathbf{R}^2$, can be separated by a line. Hence, $3 \leq \mu(\mathcal{F})$. However, the famous XOR problem is an example of a case of 4 points which can not be shattered. Hence, $\mu(\mathcal{F}) = 3$. In fact, there is no set of 4 points which can be shattered which implies $\overline{\mu}(\mathcal{F}) = 3$. Finally, if there is a set of 3 points which cannot be shattered (3 collinear points), then $\underline{\mu}(\mathcal{F}) = 2$.

2.4.2 Sontag's Classification Results. Given an artificial neural network with d-dimensional inputs, k hidden nodes with sigmoid functions, θ, on each hidden node, then F is specifically defined as described by Equation 3. Let $\mu(k, \theta, d)$ denote $\mu(\mathcal{F})$, where $\mathcal{F}$ is the set of all (k, θ)-ANNs defined by F. Similar notation will be used for $\underline{\mu}$ and $\overline{\mu}$. If direct connections are included, the notation will be $\mu^d(k, \theta, d)$. The superscript d denotes the existence of a direct connection. Again, similar notation will be used for $\underline{\mu}$ and $\overline{\mu}$.

The main results of classification capabilities in terms of established bounds on the three quantifiers will be presented here. First, consider the following two lemmas that will ease notation and provide immediate evidence of the main results. The proofs of these lemmas can be found in the reference (39:26).

Lemma 3 *For each k, θ, d, $\underline{\mu}(k, \theta, d) = \underline{\mu}(k, \theta, 1)$ and $\underline{\mu}^d(k, \theta, d) = \underline{\mu}^d(k, \theta, 1)$.*

Lemma 3 provides the independence of input dimension in the results in Theorem 4 below. That is $\underline{\mu}(k, \theta, d) = \underline{\mu}(k, \theta)$ and $\underline{\mu}^d(k, \theta, d) = \underline{\mu}^d(k, \theta)$

Lemma 4 *For any sigmoid θ, and for each k and d,*

$$\mu(k + 1, \theta, d) \geq \mu^d(k, \mathcal{H}, d)$$

and similarly for $\underline{\mu}$ and $\overline{\mu}$.

The main results are given in Theorems 4, 5, and 6 below. These theorems establish bounds on each of Sontag's V-C based mappings. Each result and proof can be found in (39).

Theorem 4 *For any sigmoid θ, and for each k,*

$$\begin{aligned} \underline{\mu}(k,\mathcal{H}) &= k+1 \\ \underline{\mu}^d(k,\mathcal{H}) &= 2k+2 \\ \underline{\mu}(k,\theta) &\geq 2k. \end{aligned}$$

Theorem 5 *For each k,*

$$\begin{aligned} 4\left\lfloor \tfrac{k}{2} \right\rfloor \leq \mu(k,\mathcal{H},2) &\leq 2k+1 \\ \mu^d(k,\mathcal{H},2) &\leq 4k+3. \end{aligned}$$

The first inequality in Theorem 5 follows from Baum (9). Actually, the bound established by Baum is for $\mu(k,\mathcal{H},d)$ (and, therefore, for $\overline{\mu}$ also) for all d.

Theorem 6 *For any sigmoid θ, and for each k,*

$$\begin{aligned} 2k+1 &\leq \overline{\mu}(k,\mathcal{H},2) \\ 4k+3 &\leq \overline{\mu}^d(k,\mathcal{H},2) \\ 4k-1 &\leq \overline{\mu}(k,\theta,2). \end{aligned}$$

Because of Lemma 4, the last statements of Theorems 4 and 6 are consequences of the previous two.

There are two interesting behaviors discovered by Sontag. Notice the apparent effect of a direct connection on the value of all of the mappings. In each case, the upper bounds approximately doubles. With the use of a standard sigmoid, the mappings exhibit similar behavior.

2.5 *Other Relevant Research*

This section is included to emphasize the on-going struggle to capture the essence of artificial neural network capabilities through the Vapnik-Chervonenkis dimension. These efforts are produced for the purposes of getting a handle on required amounts of training data and the size of the net, which appear to be intricately related. Moreover, with evolving research, a new technique is born for establishing bounds. Studying these techniques will prove useful for answering questions that arise from the ideas proposed in this dissertation. Consider the following testimonies of continued V-C dimension research.

2.5.1 V-C Dimension of Multi-Hidden-Layer ANNs. Peter L. Bartlett, has recently attacked the problem of finding lower bounds of V-C dimension for "multi-hidden-layer" ANNs (8). His results apply to the feed-forward architecture with the Heaviside function as the sigmoid function for each node and the final output of the ANN being binary. For a two hidden-layer ANN, the V-C dimension is at least equivalent to the number of connections from the input layer to other units plus one. For a three hidden-layer ANN, the results are given in the following theorem.

Theorem 7 *(8) Let M denote the set of functions represented by a three-layer, completely connected architecture with $k_0 > 0$ input units, $k_1 > 0$ first hidden-layer units, $k_2 > 0$ second hidden-layer units, and a single output unit where $k_0, k_1, k_2 \in \mathbf{N}$.*

(a) If $k_0 \geq k_1$, and $k_2 \leq 2^{k_1}/(k_1^2/2 + k_1/2 + 1)$, then

$$VC(M) \geq k_0 k_1 + k_1(k_2 - 1) + 1.$$

(b) If $1 < k_0 < k_1$, $k_2 \leq k_1$, then

$$VC(M) \geq k_0 k_1 + k_1(k_2 - 1)/2 + 1.$$

The proof of this theorem was not included in that paper. However, an explanation of his counting arguments was included. Bartlett's reasoning is based on a *defining set* for a unit, u.

Definition. A set $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbf{R}^{k_0}$ is a *defining set* for a unit, u, in a feed-forward, multi-hidden-layer, perceptron ANN with k_0-dimensional, real-valued inputs if;

- The points in X can be classified in each of $2^{\text{card}(X)}$ distinct ways by slightly perturbing the weights and threshold of unit u.
- Slightly perturbing the weights and threshold of units other than u will not affect the classification of the points in X.

Definition. A point $x \in \mathbf{R}^{k_0}$ is an *oblivious point* for the network if the classification of x is unaffected by sufficiently small perturbations of the network weights.

The next theorem lays the ground work for Bartlett's main results. It is important to note that this theorem is based on the existence of defining sets and an oblivious point.

Theorem 8 *Let M be the set of functions represented by a feed-forward, multi-hidden-layer, perceptron ANN. Consider a set of processing units U in this architecture and assume M has an oblivious point. If there is a finite defining set S_u for each unit u in U, then*

$$VC(M) \geq \sum_{u \in U} \text{card}\,(S_u) + 1.$$

Bartlett notes that the result implies that the sample size must increase at least linearly with the number of weights to guarantee the data set can be implemented. He also notes that these lower bounds hold for architectures with sigmoids. However, based on Sontag's work, there is a good chance that this lower bound for sigmoids could be tightened.

2.5.2 Hints and the V-C Dimension. In May 1993, Abu-Mostafa published a paper that used V-C dimension analysis to prove the benefits of incorporating *hints* into the learning process (2). *Hints* are known properties about the function that is being approximated by the artificial neural network. Hints reduce the class of possible functions that match the known examples. When using hints, it is appropriate to include the information in the analysis of training data requirements and ANN architecture size through V-C dimension.

The major result of this paper is that the introduction of hints does affect the V-C dimension. To show this, Abu-Mostafa defines a new quantity to represent the V-C dimension of a hint. In actuality, this is research written from the learnability side of the capabilities of ANNs, and, although it is related to the subjects in this document, the details are not directly related. The point of including the paper here is that it is yet another example of researchers using the well-established tool (V-C dimension) to show the capability benefits of a customized feature of an architecture. An additional, interesting point of Abu-Mostafa's work is that it required the customization of the definition of V-C dimension.

Abu-Mostafa's research lays additional groundwork for considering more radical alterations of the mappings to serve the purpose of customizing the capability quantifiers for specific problem sets. This is the central premise of this research outlined by this dissertation.

2.6 Conclusions

In summary, this chapter has provided enough literature review to put into perspective the problems associated with quantifying capabilities of artificial neural networks to generalize about classification data. In particular, this chapter concentrated on the feed-forward, single hidden-layer, perceptron artificial neural network. The background material was presented specifically to technically motivate the approach used in the new research in Chapters IV, V, VI and VII.

III. A Formula for Evaluating the V-C Dimension of Artificial Neural Networks

In this chapter, there is an in-depth investigation of existing proofs of evaluations of the Vapnik-Chervonenkis (V-C) dimension as stated by Baum (9) and Sontag (40). Proofs of stronger results will be given. Moreover, for fixed cardinality of a signed set, a formula for determining the required number of perceptron nodes in the hidden-layer of a feed-forward, single hidden-layer perceptron ANN is given.

Additionally, new insight into the *duality* of answering ANN capability questions are presented. The primal problem can be stated as follows: *Given an ANN architecture, what is the maximum cardinality of the classification problem which can be shattered with that architecture?* The dual problem can be stated as follows: *Given a classification problem, what is required of an ANN architecture in order to provide an implementation?* Intuitively, there would appear to be an inverse relationship. The details of this relationship are also stated in this chapter.

3.1 Clarifications and Extensions of Baum's Research

3.1.1 Baum's research revised. An unstated assumption of Baum's work (outlined in Section 2.3), is that the *chambers* of the space, $\mathbf{R}^d$, that are created by the hyperplanes associated with each hidden node can be represented by a feed-forward, single hidden-layer, perceptron ANN. Consider the function

$$F(x) = \sum_{i=1}^{k} \omega_i \mathcal{H}(v_i \cdot x - \tau_i).$$

If $\omega_i \in \{0, 1\}$, then these weights are thought of as logic variables, i.e., they indicate one side of a hyperplane or the other. Hence, every point in $\mathbf{R}^d$ (not on a hyperplane) can be identified with a unique chamber by the set of values $\{\omega_1, \omega_2, ..., \omega_k\}$.

Additionally, Baum's work, related to the primal problem, assumes the "worst case" arrangement of signed sets, geometrically, in order to prove Lemma 3. By refining "worst case" logic, stronger formulas for evaluating the V-C dimension can be achieved. Consider the following new refinement to Lemma 2 and Theorem 3. Let $P(n, \mathcal{H}, d)$ denote the minimum number of ANN hidden-layer nodes, with the Heaviside function, $\mathcal{H}$, as the sigmoid, required to guarantee that the ANN can correctly classify an arbitrary arrangement and coloring of n vectors in $\mathbf{R}^d$.

Theorem 9 *Given n vectors in $\mathbf{R}^d$ which are in general position, then*

$$P(n, \mathcal{H}, d) = \begin{cases} \left\lceil \frac{n}{d} \right\rceil & \text{for } n \text{ even} \\ \left\lceil \frac{n-1}{d} \right\rceil & \text{for } n \text{ odd.} \end{cases}$$

Proof. Let X be a finite set in $\mathbf{R}^d$ with $\text{card}(X) = n$. Let (X^+, X^-) be a dichotomy of X. Form line segments that connect each point in X^+ with its nearest neighbor in X^-, and vice versa. Note that a "worst case" arrangement of (X^+, X^-) would result in n line segments if n is even. If n is odd, then the "worst case" arrangement would result in only $n-1$ such line segments. This is due to $\text{card}(X^+) < \text{card}(X^-)$ or $\text{card}(X^-) < \text{card}(X^+)$ resulting in a redundant line segment.

The problem of implementing the dichotomy of the signed set is reduced to placing hyperplanes in the space so that each line segment is intersected. In d dimensions, a hyperplane can not be guaranteed to cut more than d line segments. Therefore, in the case n even, $\left\lceil \frac{n}{d} \right\rceil$ hyperplanes are required to intersect these line segments. In the case of n odd, $\left\lceil \frac{n-1}{d} \right\rceil$ hyperplanes are required. This provides the necessary condition. Theorem 3 gives the sufficient condition. In other words, $\left\lceil \frac{n}{d} \right\rceil$ hyperplanes are sufficient to implement the dichotomy for even n, and $\left\lceil \frac{n-1}{d} \right\rceil$ hyperplanes are sufficient for odd n. □

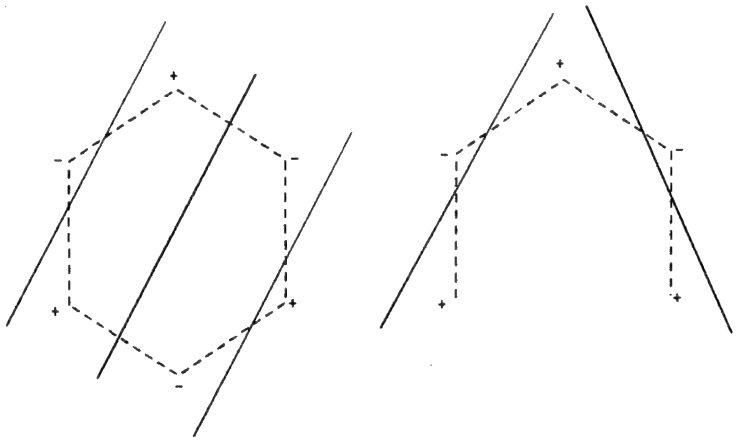


Figure 2. Even and odd ngons.

This proof is a slight modification of Baum's proof of Lemma 2 (9). See Figure 2 for a graphical view of the proof for $d = 2$. Figure 2 illustrates the difference in ANN requirements given an odd number of points versus an even number of points.

Theorem 9 has provided a closed form relation between the cardinality of signed sets to be separated and the number of processors in an ANN hidden-layer required to perform the separation. Without breaking the cases into even and odd n, equality could not be established. This seems to be a small improvement to Baum's work. However, it provided an opportunity to investigate a formula for evaluating V-C dimension.

3.1.2 Exact Values for V-C Dimension. Theorem 9 provides a value for the number of ANN *processors* required to solve an arbitrary two-class classification problem. The dual issue would be: *What is the capability of an ANN with a fixed number of hidden-layer nodes to solve two-class classification problems*? Alternatively, the question asks: *What is the value of the V-C dimension of a feed-forward, single hidden-layer, perceptron ANN?* This issue can now be addressed directly with a formula.

Let $N(p, \mathcal{H}, d)$ denote the maximum cardinality of a signed set in $\mathbf{R}^d$ independent of coloring which can be guaranteed to be separated by p processors (with associated Heaviside functions as the sigmoid) in the hidden-layer of a feed-forward, single hidden-layer, perceptron ANN. Note that $N(p, \mathcal{H}, d)$ is the V-C dimension of $\mathcal{F}_p^d$, where

$$\mathcal{F}_p^d = \{F : F(x) = \sum_{i=1}^{p} \omega_i \mathcal{H}(v_i \cdot x - \tau_i)\}$$

for $\omega_i \in \{0, 1\}$, $v_i, \in \mathbf{R}^d$, and $\tau_i \in \mathbf{R}$.

Theorem 10 *Given p processors in the hidden-layer of a feed-forward, single hidden-layer, perceptron ANN, then*

$$N(p, \mathcal{H}, d) = \begin{cases} pd + 1 & \text{for } p \text{ even, } d \in \mathbf{N} \\ pd + 1 & \text{for } p \text{ odd, } d \text{ even} \\ pd & \text{for } p \text{ odd, } d \text{ odd.} \end{cases}$$

Proof. The proof requires to simply appeal to Theorem 9 to confirm that the values can, in fact, be achieved.

Case 1 (p even, d even): Is $P(pd + 1, \mathcal{H}, d) = p$? Note that p even and d even implies that pd is even. Hence, $pd + 1$ is odd. Therefore,

$$\begin{aligned} P(pd + 1, \mathcal{H}, d) &= \left\lceil \tfrac{(pd+1)-1}{d} \right\rceil \\ &= p. \end{aligned}$$

Case 2 (p even, d odd): Is $P(pd + 1, \mathcal{H}, d) = p$? Note that p even and d odd implies that pd is even. Hence, $pd + 1$ is odd. Therefore,

$$\begin{aligned} P(pd + 1, \mathcal{H}, d) &= \left\lceil \tfrac{(pd+1)-1}{d} \right\rceil \\ &= p. \end{aligned}$$

Case 3 (p odd, d even): Is $P(pd+1, \mathcal{H}, d) = p$? Note that p odd and d even implies that pd is even. Hence, $pd+1$ is odd. Therefore,

$$\begin{aligned} P(pd+1, \mathcal{H}, d) &= \left\lceil \frac{(pd+1)-1}{d} \right\rceil \\ &= p. \end{aligned}$$

Case 4 (p odd, d odd): Is $P(pd, \mathcal{H}, d) = p$? Note that p odd and d odd implies that pd is odd. Hence, $pd+1$ is even. Therefore,

$$\begin{aligned} P(pd, \mathcal{H}, d) &= \left\lceil \frac{(pd)-1}{d} \right\rceil \\ &= \left\lceil p - \frac{1}{d} \right\rceil \\ &= p. \qquad \square \end{aligned}$$

Note that, for the particular set of ANN architectures described by $\mathcal{F}_p^d$, $N(p, \mathcal{H}, d)$ is the value of the V-C dimension of the ANN.

3.2 The Relationship of $P(n, \mathcal{H}, d)$ and $N(p, \mathcal{H}, d)$

The results given in Theorem 9 and Theorem 10 appear to present solutions to questions that are duals of each other, i.e., the results are inverses of each other. If this were the case, there would be a direct relationship between the maximum cardinality of a classification problem that can be implemented by a given number of processors and the number of processors required to *guarantee* the implementation of a given cardinality of an arbitrary arrangement of points. Intuitively, this makes sense and could be valuable when actually applying ANNs to solve a problem. However, guaranteeing a solution of an *arbitrary set* inhibits the relationship.

Consider the following. Clearly, $P(N(p, \mathcal{H}, d), \mathcal{H}, d) = p$ since this is how the results in Theorem 10 were derived. In other words, $P(n, \mathcal{H}, d)$ was evaluated at $n = N(p, \mathcal{H}, d)$. However, $N(P(n, \mathcal{H}, d), \mathcal{H}, d) \neq n$ necessarily. Consider $n = 4$ and

$d = 2$. By Theorem 9,

$$\begin{aligned} P(4, \mathcal{H}, 2) &= \left\lceil \tfrac{4}{2} \right\rceil \\ &= 2. \end{aligned}$$

However, by Theorem 10,

$$\begin{aligned} N(P(4, \mathcal{H}, 2), \mathcal{H}, 2) &= N(2, \mathcal{H}, 2) \\ &= 2 \cdot 2 + 1 \\ &= 5 \neq 4. \end{aligned}$$

This shows that P and N are not inverses as would be expected. In particular, P is the left inverse of N, i.e., $P \circ N = I$. However, $N \circ P \neq I$; i.e. P is not the right inverse of N. This conundrum originates in the semantics of the definitions of N and P. Neither of the values in the above example are incorrect. They are simply results obtained due to approaching the problem slightly differently, albeit appropriately for their purposes. Had P and N been inverses, there would have been a stronger mathematical relationship between the number of hyperplanes required to guarantee the implementation of a dichotomy and the quantified notion of ANN's capability to generalize about classification data. Consequently, the goal to build a relationship between the two questions of capability and requirements goes awry. The blame belongs to centering the quantifiers around the notion of *arbitrary* dichotomies of *arbitrary* arrangements. Hence, the stage is, once again, set to investigate a radically different approach to measuring the capabilities of ANNs.

3.3 Conclusions

In summary, this chapter has presented strengthened results of Baum's work which lead to a formula for the V-C dimension of feed-forward, single hidden-layer, perceptron artificial neural networks. Additionally, there was a discussion of the duality of ANN capability analysis and how it relates to a possible inverse relationship of the solutions. The fact that there is no complete inverse relationship serves

as additional motivation for the requirement of an approach to measuring ANNs capability to generalize about classification data other than V-C dimension based approaches.

IV. New Approach to Determining Artificial Neural Network Capabilities

There are two issues addressed in this chapter. Both are integral to the design and proper implementation of certain artificial neural network (ANN) architectures to solve classification problems. The first is an investigation of how ANN capabilities should be assessed. Note that the emphasis is not on what the capabilities are, although there should be some interesting results, but rather on designing a mapping with the specific characteristics of ANNs considered. The second issue follows from the first. Once a more informative mapping has been established, how can that mapping help determine the efficient use of ANNs to solve classification problems. In other words: *Is there any utility in using an ANN to provide a solution or does it require too many parameters to be learned from a finite training set?*

This chapter will motivate why capabilities analysis should be directed at a particular classification problem. Specifically, the quantifiers will be defined that have properties required to determine ANN capabilities. In order to evaluate these quantifiers, the complexity of the problem must be characterized along with a characterization of the capability of the sets generated from the ANN. This motivates the research given in following chapters. Ultimately, the ANN capability quantifier should be about colored arrangements of data instead of arbitrary arrangements like V-C based quantifiers. In the end, this required a complete abandonment of the V-C approach.

4.1 V-C Based Quantifiers Defined

Before defining new quantifiers, the V-C dimension based quantifiers will be defined in the new notation. This will facilitate the analysis of the pitfalls of V-C based quantifiers. Recall the following notation. Let $X \subset \mathbf{R}^d$, Y a finite subset of X, and Ω a collection of subsets of X. Let $\Delta^{\Omega}(Y)$ be the number of distinct sets

$O \cap Y$ for all $O \in \Omega$. That is,

$$\Delta^{\Omega}(Y) = \text{card}(\{O \cap Y : O \in \Omega\}).$$

Also let $\mathcal{S}_l$ denote the set of all subsets of X with cardinality l. That is, given $l \in \mathbf{N}$

$$\mathcal{S}_l = \{Y \subset X : \text{card}(Y) = l\}.$$

Now, V-C dimension and Sontag's mappings will be defined using the above notation.

Definition. Given $l \in \mathbf{N}$ and Ω defined above, let

$$\overline{m}^{\Omega}(l) = \sup\{\Delta^{\Omega}(Y) : Y \in \mathcal{S}_l\}.$$

Note that the set $\{\Delta^{\Omega}(Y) : Y \in \mathcal{S}_l\}$ is finite for each $l \in \mathbf{N}$. Hence, for all $Y \in \mathcal{S}_l$, $\Delta^{\Omega}(Y)$ is bounded above by 2^l and below by 0. Therefore, the supremum and infimum exist (are finite).

Definition. Given Ω, define

$$VC(\Omega) = \inf\{l \in \mathbf{N} : \overline{m}^{\Omega}(l) = 2^l\}$$

Definition. Given Ω, define

$$\overline{\mu}(\Omega) = \sup\{l \in \mathbf{N} : \overline{m}^{\Omega}(l) = 2^l\}$$

Note that $VC(\Omega) = \overline{\mu}(\Omega) = +\infty$ if $\overline{m}^{\Omega}(l) = 2^l$ for all $l \in \mathbf{N}$. Note that the V-C dimension and $\overline{\mu}$ are not equal. In fact, $\overline{\mu}(\Omega) = VC(\Omega) - 1$ for an arbitrary set of sets Ω. Figure 3 shows graphically the relationship between V-C dimension and $\overline{\mu}$.

Similarly, under bar symbols can be defined to yield equivalent definitions of Sontag's $\underline{\mu}$.

Definition. Given $l \in \mathbf{N}$ and Ω defined above, let

$$\underline{m}^{\Omega}(l) = \inf\{\Delta^{\Omega}(Y) : Y \in \mathcal{S}_l\}.$$

Definition. Given Ω defined above, define

$$\underline{\mu}(\Omega) = \sup\{l \in \mathbf{N} : \underline{m}^{\Omega}(l) = 2^l\}$$

Again, $\underline{\mu}(\Omega) = +\infty$ if $\underline{m}^{\Omega}(l) = 2^l$ for all $l \in \mathbf{N}$.

Now, with even more notation, a mathematical definition of μ can be stated. In order to provide this definition, it was determined that Sontag's definition of sets being *close* was actually appealing to the Hausdorff metric. The Hausdorff metric of two finite subsets of X, A and B, is defined by

$$h(A, B) = \max\{d(B, A), d(A, B)\},$$

where

$$d(A, B) = \max_{a \in A}\{\min_{b \in B} \|a - b\|\}. \quad (7{:}34)$$

Consider the following decomposition of $\mathcal{S}_l$ into disjoint subsets by defining $\mathcal{M}_l^{\Omega}$ to be the set of subsets of X with cardinality l which can be shattered by Ω, and $\mathcal{N}_l^{\Omega}$ to be the set of subsets of X with cardinality l which cannot be shattered by Ω. Specifically,

$$\mathcal{M}_l^{\Omega} = \left\{Y \in \mathcal{S}_l : \Delta^{\Omega}(Y) = 2^l\right\}$$

and

$$\mathcal{N}_l^{\Omega} = \left\{Y \in \mathcal{S}_l : \Delta^{\Omega}(Y) < 2^l\right\}.$$

Then, $\mathcal{S}_l = \mathcal{M}_l^\Omega \cup \mathcal{N}_l^\Omega$ and $\mathcal{M}_l^\Omega \cup \mathcal{N}_l^\Omega = \emptyset$ for each $l \in \mathbf{N}$.

Definition. Given Ω, define

$$\mu(\Omega) = \max\{l \in \mathbf{N} : \ \mathcal{M}_l^\Omega \text{ is dense in } \mathcal{S}_l\},$$

where $\mathcal{M}_l^\Omega$ is *dense* in $\mathcal{S}_l$ if given $\epsilon > 0$ and $S \in \mathcal{S}_l$, there exists $\widetilde{S} \in \mathcal{M}_l^\Omega$ such that $h(S, \widetilde{S}) < \epsilon$.

4.2 New Quantifiers Based on V-C Dimension

4.2.1 α Shattering. One attempt to "weaken" V-C based quantifiers (μ, $\underline{\mu}$, and $\overline{\mu}$) is to change how shattering is addressed. The first alteration to be made deals with the fact that all of the previous quantifiers are based on achieving *every* dichotomy. This is fairly restrictive. It is not always necessary to achieve every dichotomy. In this sense, the new quantifiers should require only a portion, $\alpha \in [0,1]$, of the dichotomies of a set to be achieved. Let $\alpha \in [0,1]$ and define

$$\overline{\nu}_\alpha(\Omega) = \begin{cases} +\infty & \text{if } \frac{\overline{m}^\Omega(l)}{2^l} \in [\alpha, 1] \ \forall \ l \in \mathbf{N} \\ \inf\left\{l \in \mathbf{N} : \frac{\overline{m}^\Omega(l)}{2^l} \in [0, \alpha]\right\} & \\ 0 & \text{if } \frac{\overline{m}^\Omega(l)}{2^l} \in [0, \alpha) \ \forall \ l \in \mathbf{N} \end{cases} \tag{7}$$

and

$$\underline{\nu}_\alpha(\Omega) = \begin{cases} +\infty & \text{if } \frac{m^\Omega(l)}{2^l} \in [\alpha, 1] \ \forall \ l \in \mathbf{N} \\ \sup\{l \in \mathbf{N} : \frac{m^\Omega(l)}{2^l} \in [\alpha, 1]\} & \\ 0 & \text{if } \frac{m^\Omega(l)}{2^l} \in [0, \alpha) \ \forall \ l \in \mathbf{N}. \end{cases} \tag{8}$$

Theorem 11 *The following properties are true for any Ω and $\alpha, \alpha_1, \alpha_2 \in [0,1)$.*

1. $\overline{\nu}_1 = \overline{\mu}$.
2. $\overline{\nu}_\alpha \geq \overline{\mu}$.
3. $\underline{\nu}_1 = \underline{\mu}$.

4. $\underline{\nu}_\alpha \geq \underline{\mu}$.

5. $\overline{\nu}_\alpha \geq \underline{\nu}_\alpha$.

6. If $\alpha_1 < \alpha_2$, then $\overline{\nu}_{\alpha_1} \leq \overline{\nu}_{\alpha_2}$ and $\underline{\nu}_{\alpha_1} \leq \underline{\nu}_{\alpha_2}$.

Proof. Let Ω be a collection of subsets of $X \subset \mathbf{R}^d$.

1. If $\alpha = 1$, then

$$\overline{\nu}_1(\Omega) = \begin{cases} +\infty & \text{if } \frac{\overline{m}^\Omega(l)}{2^l} = 1 \ \forall\, l \in \mathbf{N} \\ \inf\left\{l \in \mathbf{N} : \frac{\overline{m}^\Omega(l)}{2^l} \in [0,1]\right\} \end{cases}$$

   which is $\overline{\mu}(\Omega)$.

2. Since

$$\inf\{l \in \mathbf{N} : \frac{\overline{m}^\Omega(l)}{2^l} \in [0,\alpha]\} \geq \sup\{l \in \mathbf{N} : \overline{m}^\Omega(l) = 2^l\}$$

   then $\overline{\nu}_\alpha \geq \overline{\mu}$.

3. Follows similarly to $\overline{\nu}$.

4. Follows similarly to $\overline{\nu}$.

5. By definition of $\overline{m}^\Omega(l)$ and $\underline{m}^\Omega(l)$.

6. Let $\alpha_1 < \alpha_2$, then $[0,\alpha_1] \subset [0,\alpha_2]$. Hence,

$$\inf\{l \in \mathbf{N} : \frac{\overline{m}^\Omega(l)}{2^l} \in [0,\alpha_1]\} \leq \inf\{l \in \mathbf{N} : \frac{\overline{m}^\Omega(l)}{2^l} \in [0,\alpha_2]\}.$$

   Therefore, $\overline{\nu}_{\alpha_1} \leq \overline{\nu}_{\alpha_2}$. Proof of $\underline{\nu}_{\alpha_1} \leq \underline{\nu}_{\alpha_2}$ is similar. □

See Figures 3 and 4. These figures illustrate the relationship between V-C based quantifiers and the new quantifiers.

For the "ν_α-equivalent" to μ, more notation is needed.

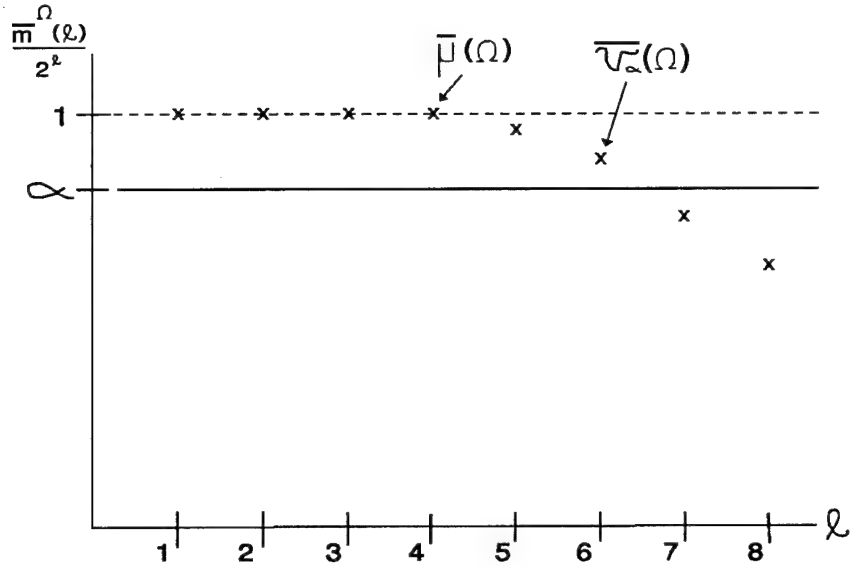


Figure 3. Relationship of $\overline{\mu}$ and $\overline{\nu}_\alpha$.

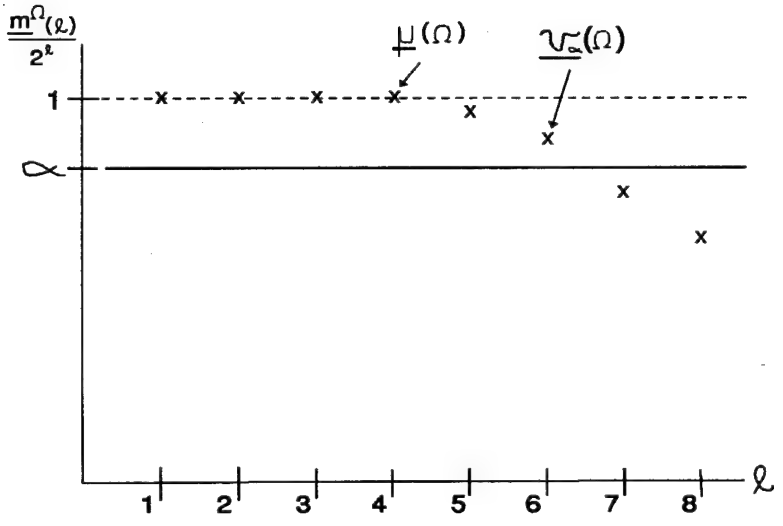


Figure 4. Relationship of $\underline{\mu}$ and $\underline{\nu}_\alpha$.

Definition. Given $l \in \mathbf{N}$ and $\epsilon > 0$, define the ϵ-ball centered at the set $S \in \mathcal{S}_l$ as

$$B_l(S, \epsilon) = \{T \in \mathcal{S}_l : h(T, S) < \epsilon\}.$$

Hence, $B_l(S, \epsilon)$ is the set of sets of cardinality l that are within ϵ to S with respect to the Hausdorff metric. Now the definition of $\widetilde{\nu}_{\epsilon,\alpha}(\Omega)$ can be defined.

Definition. Given $l \in \mathbf{N}$ and Ω, $B_l(S, \epsilon)$ defined above, let

$$\widetilde{m}_{\epsilon}^{\Omega}(l) = \inf\{\Delta^{\Omega}(T) : T \in B_l(S, \epsilon)\}.$$

Now, given ϵ and $\alpha \in [0, 1]$, define

$$\widetilde{\nu}_{\epsilon,\alpha}(\Omega) = \begin{cases} +\infty & \text{if } \frac{\widetilde{m}_{\epsilon}^{\Omega}(l)}{2^l} \in [\alpha, 1] \ \forall\, l \in \mathbf{N} \\ \inf\left\{l \in \mathbf{N} : \frac{\overline{m}^{\Omega}(l)}{2^l} \in [0, \alpha]\right\} & \\ 0 & \text{if } \frac{\widetilde{m}_{\epsilon}^{\Omega}(l)}{2^l} \in [0, \alpha) \ \forall\, l \in \mathbf{N} \end{cases} \tag{9}$$

4.2.2 Between "At Least One" and "Every" Set. Another deviation from the base set of quantifiers should be about the rigid difference between $\underline{\mu}$ and $\overline{\mu}$ and, likewise, between $\underline{\nu}_{\alpha}$ and $\overline{\nu}_{\alpha}$. For $\underline{\nu}_{\alpha}$ to have the value l, it is required that *every* set of cardinality l must satisfy the conditions prescribed. For $\overline{\nu}_{\alpha}$ to have the value l, it is required that *only one* set of cardinality l must satisfy the prescribed conditions. This is an extreme difference. In an attempt to "weaken" the mapping, another parameter is incorporated. Define $\mathcal{S}_l^{\beta}$ as some "reduced portion" of $\mathcal{S}_l$, i.e., $\mathcal{S}_l^{\beta} \subset \mathcal{S}_l$. Then, redefine $\overline{m}_{\beta}^{\Omega}(l)$ and $\overline{\nu}_{\alpha}$ on the reduced set $\mathcal{S}_l^{\beta}$. Consider

$$\overline{m}_{\beta}^{\Omega}(l) = \sup\{\Delta^{\Omega}(Y) : Y \in \mathcal{S}_l^{\beta}\}$$

and

$$\underline{m}_{\beta}^{\Omega}(l) = \inf\{\Delta^{\Omega}(Y) : Y \in \mathcal{S}_l^{\beta}\}.$$

Now, define

$$\overline{\nu}_{\alpha,\beta}(\Omega) = \begin{cases} +\infty & \text{if } \frac{\overline{m}_{\beta}^{\Omega}(l)}{2^l} \in [\alpha, 1] \ \forall\, l \in \mathbf{N} \\ \inf\left\{ l \in \mathbf{N} : \frac{\overline{m}_{\beta}^{\Omega}(l)}{2^l} \in [0, \alpha] \right\} & \\ 0 & \text{if } \frac{\overline{m}_{\beta}^{\Omega}(l)}{2^l} \in [0, \alpha) \ \forall\, l \in \mathbf{N} \end{cases} \tag{10}$$

and

$$\underline{\nu}_{\alpha,\beta}(\Omega) = \begin{cases} +\infty & \text{if } \frac{\underline{m}_{\beta}^{\Omega}(l)}{2^l} \in [\alpha, 1] \ \forall\, l \in \mathbf{N} \\ \sup\{ l \in \mathbf{N} : \frac{\underline{m}_{\beta}^{\Omega}(l)}{2^l} \in [\alpha, 1] \} & \\ 0 & \text{if } \frac{\underline{m}_{\beta}^{\Omega}(l)}{2^l} \in [0, \alpha) \ \forall\, l \in \mathbf{N}. \end{cases} \tag{11}$$

Note that, in a sense, μ and $\tilde{\nu}$ already have this feature since sets not in general position are not considered. So, the β parameter just extends that freedom to sets with no particular characteristics and, hence, a version of $\tilde{\nu}$ incorporating β is redundant.

4.2.3 Inadequacies of $\overline{\nu}_{\alpha,\beta}$ and $\underline{\nu}_{\alpha,\beta}$. The above attempts to customize V-C based quantifiers for ANN capabilities analysis were attempts to incorporate the desired properties at the wrong point of analysis. They are attempts to characterize separating surfaces when, in fact, this customization needs to be transferred to the problem set.

Additionally, it is clear that there are two distinct, but obviously related, questions to be answered: *What are the particular requirements of an ANN architecture in order to solve a given classification problem?* and *How do different ANN architectures compare in capability?* Consider the V-C based quantifiers. Theoretically, they provide *worst case* notions of capabilities. Chapter III provided theorems for inverting these concepts to get *worst case* requirements. How useful is this since it is the worst case arrangements of data being considered? The crux of the new research

in this dissertation is that it would be helpful to perform similar analysis on the colored arrangements in a way that is accessible directly for implementation.

4.3 Conclusions

In summary, it has been shown that the difficulty of measuring ANNs' capability is partially because each approach discussed here has sought to assign a value to the capability. Alternatively, consider ordering a set of ANNs based on capability. This would provide a means of comparing ANNs based on capability without having to assign a value. Additionally, it should be clear that the comparison should be based on an ANNs' ability to solve specific classification problems not arbitrary classification problems.

V. Artificial Neural Networks, Combinatorial Geometry, and Lattice Theory

This chapter introduces concepts of combinatorial geometry of hyperplane arrangements and lattice theory that are relevant to studying capabilities of artificial neural networks (ANNs). These concepts provide the basis for a generalized approach (derived in Chapter VI) to quantifying ANNs' capabilities based on *invariants.* Roughly speaking, invariants are mathematical objects which do not change given certain prescribed transformations of their domain. These invariants impinge upon many aspects of the geometry of the arrangement. Borrowing from these ideas, ANN capability quantifiers will be constructed in Chapter VII. To analyze some invariants an underlying lattice structure is required. Therefore the lattice structures of ANNs will be defined in this chapter. Finally, it will be shown that the V-C dimension of a feed-forward, single hidden-layer, perceptron ANN is equivalent to the number of *chambers* of the hyperplane arrangement defined by the ANN.

A majority of the results of combinatorial geometry relies on a lattice structure. Hence, this chapter will investigate the lattice structures of feed-forward, single hidden-layer, perceptron artificial neural networks. Additionally, lattice theory provides a mechanism for comparing sets. Hence, establishing a lattice on sets generated ANNs will give insight into how architectures compare in their ability to generalize.

Although the study of hyperplane arrangements is a relatively young field of mathematics (important works by B. Grünbaum published in 1971 (21) and by Zaslavsky in 1975 (50)), there is a wealth of theorems and algorithms answering current applied mathematics problems. In particular, methods for counting geometric structures of an arrangement such as edges, faces, and chambers have been established (3). The importance of these counting algorithms for measuring ANN capabilities is discussed in this chapter.

The link between the study of arrangements of hyperplanes and the Vapnik-Chervonenkis (V-C) dimension relies on *chamber counting.* Chamber counting is a basic objective of many combinatorial geometric algorithms. One example of counting is based upon the method of deletion-restriction which relies on the Poincaré polynomial of an arrangement of hyperplanes (36). This notion is intricately related to quantifying the capabilities of ANNs.

It is important to note that results obtained here are exclusive to the ANN architecture described in Chapter I. Also, all combinatorial geometric results in this chapter are based on the assumption that the arrangement of hyperplanes is in general position. That is, if any two planes have a common line, then the line is distinct, and if any three planes have a common point, then the point is distinct.

5.1 Combinatorial Geometry - The Basics

In a broad sense, combinatorial geometry is the theory of arithmetic invariants of finite sets of points in projective space (50). Some important invariants depend upon the lattice structure of the subspaces formed by a point set. Of particular importance to ANN research, are point sets which are the duals of an arrangement of hyperplanes. Fortunately, lattice structures can be defined for arrangements of hyperplanes (50), making the rich results about *geometric lattices* available for ANN analysis. Geometric lattices are lattices that satisfy additional conditions requiring that the arrangement be *central*, that is, each hyperplane contains a point common to all of the hyperplanes in the arrangement (1).

5.1.1 Lattice Structures. Lattice theory is required to exploit the results of combinatorial geometry for ANN analysis. Lattice theory, in general, is the study of orderings. An ordering is a binary relation, $\preceq$, which can be read as "is contained in", "is a part of", or "is less than or equal". The purpose of an ordering is to be able to compare elements of a set. To make meaningful comparisons, an ordering is

required to satisfy certain properties. The number and type of properties required dictates directly what can be ascertained from the comparison of elements in the set.

The most basic properties define a *partially ordered set.*

Definition. A *partially ordered set* is a set, P, together with a binary relation, $\preceq$, which satisfies, for all $x, y, z \in P$, the following properties:

1. $x \preceq x$. (Reflexive)
2. If $x \preceq y$ and $y \preceq x$, then $x = y$. (Antisymmetry)
3. If $x \preceq y$ and $y \preceq z$, then $x \preceq z$. (Transitive)

If $x \preceq y$ and $x \neq y$, then we write $x \prec y$. Note that the term *partially* is used to indicate that the relation $\preceq$ is not necessarily closed. In other words, not all elements of the set are required to be comparable. Hence, there may be $x, y \in P$ for which $x \neq y$, $x \not\preceq y$, $y \not\preceq x$. Additionally, y is said to *cover* x if $x \prec y$ and there is no $z \in P$ such that $x \prec z \prec y$. Also, if there is a unique element $z \in P$ such that $z \preceq x$ for all $x \in P$, then z is called the *zero element* of P.

Definition. Let $(P, \preceq)$ be a partially ordered set. A *lower bound* (*upper bound*) of a subset X of P, is an element $a \in P$ ($b \in P$) such that $a \preceq x$ for all $x \in X$ ($x \preceq b$ for all $x \in X$).

Definition. Let $(P, \preceq)$ be a partially ordered set. A *greatest lower bound* (*least upper bound*) of X is an lower bound, $\tilde{a} \in P$, (upper bound $\tilde{b} \in P$) such that $a \preceq \tilde{a}$ ($\tilde{b} \preceq b$) for all lower bounds (upper bounds) of X.

Definition. Let $(P, \preceq)$ be a partially ordered set. Let $x_1, x_2, ..., x_n \in P$ such that $x_1 \prec x_2 \prec ... \prec x_n$, then $x_1 \prec x_2 \prec ... \prec x_n$ is said to be a *chain.* A *saturated chain* is a chain, $x_1 \prec x_2 \prec ... \prec x_n$, such that x_{i+1} covers x_i for all $i < n$. The *length* of a chain is defined as one less than its cardinality.(5:14)

Definition. Let $(P, \preceq)$ be a partially ordered set. The *rank function* of an element $x \in P$, $r(x)$ is defined as the maximum length of all saturated chains from z to x where $z = \text{glb}(P)$.(5:14)

Definition. Let $(P, \preceq)$ be a partially ordered set. The *meet* of any two elements $x, y \in P$, denoted by $x \wedge y$, is defined as the greatest lower bound of the set $\{x, y\}$. That is

$$x \wedge y = \text{glb}\{x, y\}.$$

The *join*, denoted by $x \vee y$, as the least upper bound of the set $\{x, y\}$. That is

$$x \vee y = \text{lub}\{x, y\}.$$

Note that the definitions of the meet and join are inherently dependent on the ordering. In other words, the ordering defines the meet and join.

Definition. A partially ordered set, $(P, \preceq)$, with operations meet, $\wedge$, and join, $\vee$, such that P is closed with respect to $\wedge$ and $\vee$ is called a *lattice* and is denoted $(P, \preceq, \wedge, \vee)$.

One should note that it is *not* always the case that the set, P, is closed with respect to the meet and join operations. Hence, we have the following definitions.

Definition. A partially ordered set, $(P, \preceq)$, with the operation meet, $\wedge$, such that P is closed with respect to $\wedge$ is called a *meet semi-lattice* and is denoted $(P, \preceq, \wedge)$.

Definition. A partially ordered set, $(P, \preceq)$, with the operation join, $\vee$, such that P is closed with respect to $\vee$ is called a *join semi-lattice* and is denoted $(P, \preceq, \vee)$.

The term semi-lattice will be used to refer to a join semi-lattice or a meet semi-lattice.

Now, consider the properties of a lattice as presented by the following lemma.

Lemma 5 *Let $(P, \preceq, \wedge, \vee)$ be a lattice, then, for all $x, y, z \in P$, the following laws hold true:*

(1) $x \wedge x = x,\ x \vee x = x$. *(Idempotent)*

(2) $x \wedge y = y \wedge x,\ x \vee y = y \vee x$. *(Commutative)*

(3) $x \wedge (y \wedge z) = (x \wedge y) \wedge z,\ x \vee (y \vee z) = (x \vee y) \vee z$. *(Associative)*

(4) $x \wedge (x \vee y) = x \vee (x \wedge y) = x$. *(Absorption)*

Moreover,

(5) $x \preceq y \iff x \wedge y = x$ *and* $x \preceq y \iff x \vee y = y$. *(Consistency) (12)*

With the structures described above, a large variety of *combinatorial problems,* such as the chamber counting problem, can be addressed. Combinatorial problems can often be restated as *critical problems* (17). Critical problems bridge lattice theory to combinatorial geometry of hyperplane arrangements. Critical problems can be solved by analyzing the characteristic polynomial, also known as the Poincaré polynomial, of a lattice or semi-lattice.

It will be shown by the new research in this dissertation that hyperplane arrangements can have a lattice structure and, in general, have a semi-lattice structure. Moreover, the number of chambers created by an arrangement can be ascertained from the characteristic polynomial defined on the lattice (or semi-lattice) of the hyperplane arrangement. The novelty of this approach to solving geometric problems is that the solutions are arithmetic invariants described by the characteristic polynomial and depend only on the geometry of the original point set (17).

5.2 The Theory of Arrangements of Hyperplanes

Of particular interest to the research in this dissertation are the results of combinatorial geometry applied to arrangements of hyperplanes. This is yet another relatively new area of rich mathematics that is pertinent to the study of ANNs.

Consider a finite set of hyperplanes (translated subspaces with dimension $d-1$) of a Euclidean or projective d-dimensional space which will be referred to as an *arrangement.* When these hyperplanes are removed, the remainder of the space

is partitioned into disjoint subsets known as *chambers*, each one a d-dimensional polyhedron (not necessarily bounded). The arrangement is said to *partition the space by hyperplanes.* How many chambers are created by the partition? How many vertices result from the hyperplane intersections? Are two arrangements equivalent geometrically? These are some of the questions that the theory of arrangements of hyperplanes seeks to answer and will be a central concept in the development of geometric measures of ANN capabilities.

5.2.1 The Cut-Intersection Semi-Lattice. Let A denote an arrangement of hyperplanes (so $l \in A$ is a particular hyperplane). The cut-intersection semi-lattice of an arrangement is actually defined on the intersections of the hyperplanes and is denoted $L(A)$. That is,

$$L(A) = \left\{ \bigcap_{l \in T} l : \bigcap_{l \in T} l \neq \emptyset \text{ for all } T \subseteq A \right\}.$$

An element, $x \in L(A)$, could be any k-dimensional translated subspace, for $k \in \{0, 1, 2, ..., d-1\}$, contained in some hyperplane $a \in A$. The partial ordering, $\preceq$, is chosen to be reverse set containment. That is, for $x, y \in L(A)$, write $x \preceq y$ if and only if $x \supseteq y$. Then, $(L(A), \preceq)$ is a partially ordered set. (50)

The ordering is chosen so that the minimal element of $L(A)$ is the entire space containing the hyperplanes. The existence of this minimal element is important for following results. Note that there is no guarantee of the existence of a *maximal* element which is expected to be the empty set. Recall that $\emptyset \notin L(A)$. However, with stronger conditions on the arrangement A, the existence of an maximal element can be guaranteed.

Now, consider the definitions of meet and join for $L(A)$. since the ordering is *reverse* set inclusion, the definitions are not intuitive. In fact, the meet operation, as prescribed by the definitions in Section 5.1.1, is not a closed operation. Hence,

$L(A)$ will be called a semi-lattice. A modified definition of the meet is given so that the set is closed with respect to the meet. Define the modified meet operation of $x, y \in L(A)$ by

$$x \wedge y = \text{glb}\{x, y\} = \cap\{l \in A : l \supseteq x \cup y\}.$$

The join operation (as defined by the definition in Section 5.1.1) of $x, y \in L(A)$ becomes

$$x \vee y = \text{lub}\{x, y\} = x \cap y.$$

Note that $x \cap y$ is not necessarily in $L(A)$, since $x \cap y$ could be empty. Hence $L(A)$ is not closed with respect to the join operation. It has been shown that $(L(A), \preceq, \wedge, \vee)$ satisfies all of the properties of Lemma 5 (50). Hence, $(L(A), \preceq, \wedge, \vee)$ will be referred to as the cut-intersection semi-lattice of an arrangement A. Note that it is a *semi-*lattice since $L(A)$ is not closed with respect to the join operation.

Finally, consider the lattice structure of an arrangement that is said to be *centered.*

Definition. An arrangement, A, is *centered* if

$$\bigcap_{l \in A} l \neq \emptyset.$$

Note that $L(A)$ defined on a central arrangement, A, is assured to have maximal element. In fact, $(L(A), \preceq, \wedge, \vee)$, is a lattice (with the prescribed meet operation, not the modified meet operation) and is often referred to as a *geometric lattice.* (36:24)

5.3 Chamber Counting and the Poincaré Polynomial

Chamber counting is a fundamental combinatorial geometry problem. Not only does chamber counting provide valuable insight into arrangements (for the purposes of this research) of hyperplanes, but it also serves as a benchmark problem for testing algorithms or methods much like the map coloring problem. In fact, there is

a significant amount of research dedicated to counting methods for many geometric entities such as chambers, faces, vertices, and edges.

In 1889, S. Roberts published breakthrough research that included a very simple formula for counting chambers. According to Roberts, the number of regions formed by an arbitrary arrangement of n lines in the Euclidean plane is equal to the number of regions formed by n lines in general position, minus the number of regions lost because of multiple points, minus the number of regions lost because of parallel lines. This formula led to various algorithms for determining actual counts, one of which was proposed by T. Zaslavsky (50) in 1975. Zaslavsky's approach is known as the method of deletion and restriction. In addition, he showed that the recursive formula used for the method of deletion and restriction produces the same result for counting chambers as the characteristic polynomial evaluated at one.

5.3.1 The Method of Deletion and Restriction. Let A be an arrangement in the Euclidean d-dimensional space, and let $l \in A$ be a hyperplane.

Definition. The *deleted arrangement about* l, is defined as $A' = A \backslash \{l\}$. The *restricted arrangement about* l is defined as $A'' = \{K \cap l \mid K \in A'\}$.

The triple (A, A', A'') can be used to solve the problem of counting *chambers*. Let $C(A)$ denote the set of chambers formed by A. Zaslavsky showed that

$$\text{card}(C(A)) = \text{card}(C(A')) + \text{card}(C(A'')).$$

To prove this recursion, let P be the set of chambers in $C(A')$ that intersects the distinguished hyperplane, l, and let Q be the set of chambers in $C(A')$ that does not intersect l. Obviously, $\text{card}(C(A')) = \text{card}(P) + \text{card}(Q)$. Note that the hyperplane l divides each chamber of P into 2 chambers and does not intersect the chambers of Q. Hence, $\text{card}(C(A)) = 2\text{card}(P) + \text{card}(Q)$. In fact, there is a bijection between P

and $C(A'')$ given by $C \longrightarrow C \cap l$. Therefore, $\text{card}(C(A'')) = \text{card}(P)$, which provides the recursion. (50) (36)

5.3.2 The Poincaré Polynomial. Zaslavsky also showed that a similar recursion holds for the *characteristic polynomial.* The characteristic polynomial is a specific instantiation of the *Poincaré polynomial,* which is a polynomial defined on the geometric structure of an arrangement and is used for the analysis of invariants about the cut-intersection semi-lattice, $L(A)$. The Poincaré polynomial is one of the most important combinatorial invariants of an arrangement. Its properties provide insight into the structures of an arrangement. The Poincaré polynomial is defined based on the Möbius function and a rank function.

The Möbius function, defined on $L(A)$, is a binary mapping $\mu : L(A) \times L(A) \rightarrow \mathbf{Z}$. It provides a characterization for arrangement density, by characterizing the relationships of the subsets based on the partial ordering defined for $L(A)$. In general, there is not an explicit formula for μ. However, for fixed x, the values of $\mu(x, y)$ may be computed recursively as follows. For $x, y, z \in L(A)$,

$$\mu(x,y) = \begin{cases} 1 & \text{if} \quad x = y \\ -\sum\limits_{x \preceq y \preceq z} \mu(x,z) & \text{if} \quad x \preceq y \\ 0 & \text{else.} \end{cases} \tag{12}$$

Note that if a function, Ψ, satisfies the above properties, then $\Psi = \mu$. (36:33). That is, μ is uniquely defined. The rank function, defined on $L(A)$, is as defined in Section 5.1.1. Note that in the case of $L(A)$, the rank function on $x \in L(A)$ can be defined as $r(x) = co\dim(x)$ (36:24).

With the Möbius function and the rank function, the characteristic polynomial, π, can be defined. Let A be an arrangement, and let $L(A)$ be the cut-intersection semi-lattice defined on A. Define $\mu(x) = \mu(V, x)$, where V represents the entire space and, hence, is the greatest lower bound of $L(A)$ since $L(A)$ is ordered by reverse

inclusion. Then, the characteristic polynomial of A is defined in (36) as

$$\pi(A,t) = \sum_{x \in L(A)} \mu(x)(-t)^{r(x)}. \tag{13}$$

Note that for the special empty arrangement, $A = \emptyset$, then, by definition, $\pi(A,t) = 1$. Zaslavsky also showed that given an arrangement A and $a \in A$, then the following recursion formula for the characteristic polynomial holds:

$$\pi(A,t) = \pi(A',t) + t\pi(A'',t).$$

Then, since $\text{card}(C(A))$ and $\pi(A,1)$ have the same value for $A = \emptyset$ and satisfy the same recursion for deletion and restriction, it is true, by uniqueness, that

$$\text{card}(C(A)) = \pi(A,1).$$

This is an important result that led to more analytical approaches for counting chambers and other geometric entities. (50) (36)

5.4 The Lattice Structure and Combinatorial Geometry of Artificial Neural Networks

Sections 5.1-5.3 defined a lattice, the specifics of the cut-intersection semilattice defined on the set of intersections of hyperplanes, and the analysis that can be accomplished through the Poincaré Polynomial. In parallel, this section provides the details of the same structures defined on the set of sets derived from a feedforward, single hidden-layer, perceptron artificial neural network.

The reason for establishing a lattice structure on which to perform ANN capability analysis is two-fold. The first is the general notion that, ultimately, an ordering on ANN architectures is sought which is based on capability defined on invariants. Well-behaved orderings are one of the outcomes of lattice theory along with provid-

ing a basis from which invariants can be extracted. Obtaining those invariants is the second reason a lattice is sought since invariants will be used to characterize the *complexity* of signed sets.

5.4.1 The Cut-Intersection Semi-Lattice of ANNs. Given a fixed feed-forward, single hidden-layer, perceptron ANN, there is an implied arrangement of hyperplanes which directly establishes a finite set of half-spaces. The intersection of the half-spaces is accomplished through the logic layer of the network. The result is a set of sets. Therefore, with set containment providing an ordering (by reverse inclusion), intersection as the meet operation and union as the join, this is a specific example of the cut-intersection semi-lattice defined in Section 5.2.

First, notation will be established. Consider the fixed ANN, (ω_i, v_i, τ_i, k fixed),

$$F(x) = \sum_{i=1}^{k} \omega_i \mathcal{H}(v_i \cdot x - \tau_i), \tag{14}$$

where $x \in \mathbf{R}^d$ is the input vector, $\mathcal{H}$ is the Heaviside function, $v_i \in \mathbf{R}^d$ for $i = 1, \ldots, k$ are the weight vectors, $\tau_i \in \mathbf{R}$ for $i = 1, 2, ..., k$ are the threshold values at each of the k nodes and $\omega_i \in \{0, 1\}$ for $i = 1, \ldots, k$ are known values. This defines an instantiation of an architecture, which corresponds to a finite arrangement of hyperplanes, $A = \{l_1, l_2, l_3, ..., l_k\}$, where

$$l_i = \{x \in \mathbf{R}^d : \ L_i(x) = v_i \cdot x - \tau_i = 0\}$$

and L_i defines an affine mapping $\mathbf{R}^d \to \mathbf{R}$. That is, each L_i corresponds to a $(d-1)$-dimensional hyperplane, $l_i = \{x \in \mathbf{R}^d : L_i(x) = 0\}$. Assume that each of the k hyperplanes are distinct. Since each hyperplane naturally divides $\mathbf{R}^d$ in half, define the corresponding set of half-spaces $H = \{h_1, h_2, h_3, ..., h_k\}$ as

$$h_i = \{x \mid v_i \cdot x - \tau_i > 0\} \subset \mathbf{R}^d.$$

Note that by switching the sign of v_i and τ_i, h_i represents the "other half" of the space or the interior of the complement of the half-space. Define

$$h_i^c = \{x \mid v_i \cdot x - \tau_i < 0\} \subset \mathbf{R}^d.$$

Note that this is the *interior* of the complement of h_i, not exactly the complement.

Additionally, consider the set of all possible arrangements of hyperplanes in $\mathbf{R}^d$. Define

$$\mathbf{A}^d = \{A = \{l_1, l_2, ..., l_k\} : k \in \mathbf{N},\ l_i = \{x : L_i(x) = 0\} \text{ for each } i = 1, 2, ..., k\}.$$

Note that by definition, any $A \in \mathbf{A}^d$ can be defined by equation 14. Also, define the corresponding set of sets of half-spaces, $\mathbf{H}^d$

$$\mathbf{H}^d = \{H = \{h_1, h_2, h_3, ..., h_k\} : k \in \mathbf{N}\}.$$

Note that, $\mathbf{A}^d$ and $\mathbf{H}^d$ can be partitioned dependent on the type of architectures that are being investigated for capability analysis. For example, a partition based on the number of hyperplanes, k, in an arrangement gives a simple organization of all architectures. Specifically, for each $\kappa \in \mathbf{N}$, let

$$\mathcal{F}_\kappa^d = \{A = \{l_1, l_2, l_3, ..., l_\kappa\} \in \mathbf{A}^d\}.$$

Note that $\mathcal{F}_\kappa^d$ can be thought of as the set of k hyperplane arrangements generated from a feed-forward, single hidden-layer, perceptron ANN with κ nodes in the hidden-layer and d-dimensional input since each node corresponds to a hyperplane. Moreover,

$$\mathbf{A}^d = \bigcup_{\kappa \in \mathbf{N}} \mathcal{F}_\kappa^d.$$

Hence, $\mathbf{A}^d$ is the set of all feed-forward, single hidden-layer, perceptron ANN with an arbitrary set of nodes in the hidden-layer and d-dimensional input.

Now, consider the finite set of all intersections of the hyperplanes in an arrangement, A

$$\mathcal{C}_A \equiv \left\{ \bigcap_{l \in T} l \ : \ \bigcap_{l \in T} l \neq \emptyset \text{ for all } T \subseteq A \right\}.$$

Define the partial ordering on $\mathcal{C}_A$, $\preceq$, as reverse set inclusion. In other words, let x, $y \in \mathcal{C}_A$, then $x \preceq y$, if and only if $y \subseteq x$.

Theorem 12 *Given an arrangement, $A \in \mathbf{A}^d$, corresponding cut-intersection set, $\mathcal{C}_A$, and ordering, $\preceq$, then $(\mathcal{C}_A, \preceq)$ is a partially ordered set.*

Proof. Let x, y, $z \in \mathcal{C}_A$. Clearly, for all $x \in \mathcal{C}_A$, $x \preceq x$ since $x \subseteq x$. Hence, $\preceq$ is reflexive. Assume $x \preceq y$ and $y \preceq x$. This implies that $x \subseteq y$ and $y \subseteq x$. Hence, $x = y$. Therefore, $\preceq$ is antisymmetric. Assume $x \preceq y$ and $y \preceq z$. This implies that $x \subseteq y$ and $y \subseteq z$, which implies $x \subseteq z$. Hence, $x \preceq z$ implying that $\preceq$ is transitive.

Therefore, combining implies $(\mathcal{C}_A, \preceq)$ is a partially ordered set. □

Now, define the join as

$$x \vee y = \text{lub}\{x, y\} = x \cap y$$

and the meet as

$$x \wedge y = \text{glb}\{x, y\} = \cap\{l \in A : l \supseteq x \cup y\}.$$

Theorem 13 *Given an arrangement, $A \in \mathbf{A}^d$, the set $\mathcal{C}_A$, the ordering $\preceq$, and the operations $\vee$, and $\wedge$, then $(\mathcal{C}_A, \preceq, \vee, \wedge)$ is a cut-intersection semi-lattice.*

Proof. $(\mathcal{C}_A, \preceq, \vee, \wedge)$ is a specific example of the cut-intersection semi-lattice defined in general in Section 4.2. □

From here, refer to $(\mathcal{C}_A, \preceq, \vee, \wedge)$ as the cut-intersection semi-lattice of a feed-forward, single hidden-layer, perceptron artificial neural network.

5.4.2 The Characteristic Polynomial of $(\mathcal{C}_A, \preceq, \vee, \wedge)$. With the cut-intersection semi-lattice, $(\mathcal{C}_A, \preceq, \vee, \wedge)$, defined, the combinatorial geometry of the hyperplane arrangement can be investigated. In particular, the characteristic polynomial can be defined.

Consider the Möbius function defined as

$$\mu(x,y) = \begin{cases} 1 & \text{if} \quad x = y \text{ and } x \in \mathcal{C}_A \\ -\sum\limits_{x \preceq z \preceq y} \mu(x,z) & \text{if} \quad x,y,z \in \mathcal{C}_A \text{ and } x \preceq y \\ 0 & \text{else.} \end{cases} \tag{15}$$

Again, for convenience, define $\mu(x) = \mu(V, x)$, where V is the greatest lower bound of $\mathcal{C}_A$ which would be $\mathbf{R}^d$. Let $r(x) = co\dim(x)$. Then, the characteristic polynomial of any $A \in \mathbf{A}^d$ can be defined as

$$\pi(A,t) = \sum_{x \in \mathcal{C}_A} \mu(x)(-t)^{r(x)}. \tag{16}$$

Now, structure is in place to analytically and deterministically evaluate certain invariants about an ANN architecture.

5.4.3 The Relationship Between the Poincaré Polynomial and the Vapnik-Chervonenkis Dimension. What does the Poincaré polynomial have to do with artificial neural networks? First, the fact that Zaslavsky introduced an analytical method for determining the value of $\text{card}(C(A))$ has made the venture into the study of combinatorial geometry of arrangements of hyperplanes worthwhile. Chamber cardinality, through Zaslavsky's work, is now an accessible geometric invariant. In fact, it is now pertinent to consider the relationship of combinatorial geometry and

the arrangement of hyperplanes that result from the ANN architecture discussed in Chapter I.

By Zaslavsky's result described above, for each $A \in \mathcal{F}_\kappa^d$, there is a lattice $L(A)$ and corresponding Möbius and rank function to define the characteristic polynomial, π. Moreover, for each $A \in \mathcal{F}_\kappa^d$, $\text{card}(C(A)) = \pi(A, 1)$.

Let $A^* \in \mathcal{F}_\kappa^d$ be an arrangement of hyperplanes such that the maximum number of chambers is achieved. This happens when the arrangement is in general position and $\text{card}(A^*) = \kappa$. (Recall that an arrangement is in general position if, when any two planes have a common line, the line is distinct; and when any three planes have a common point, the point is distinct.) More formally,

$$\text{card}(C(A^*)) = \sup\left\{\text{card}(C(A)) : A \in \mathcal{F}_\kappa^d\right\}.$$

Define $L(A^*)$ as before with the reverse set inclusion ordering. Note that since A^* is just an instance of an arbitrary arrangement of hyperplanes, then $L(A^*)$ is a cut-intersection semi-lattice. Therefore, it can substantiate the Möbius function, μ, the rank function, r, and the characteristic polynomial, π.

Theorem 14 *Let $A^* \in \mathcal{F}_\kappa^d$, be an arrangement of hyperplanes defined by a trained ANN architecture such that $\text{card}(C(A^*)) = \sup\left\{\text{card}(C(A)) : A \in \mathcal{F}_\kappa^d\right\} = \kappa$. Assume A^* is in general position. Then,*

$$VC(\mathcal{F}_\kappa^d) = \text{card}(C(A^*)).$$

Proof. Let $n = \text{card}(C(A^*))$. Choose a set, $X \in \mathbf{R}^d$, such that $\text{card}(X) = n$ and each $x \in X$ lies in a different chamber of A^*. Note, since the points are all separated from each other in the chambers, then the hyperplanes of A^* that form the chambers can implement any dichotomy of X. This demonstrates a set of

n points which can be shattered by the arrangement, A^*, and since $A^* \in \mathcal{F}_\kappa^d$, then $VC(\mathcal{F}_\kappa^d) \geq n$.

In order to show that $VC(\mathcal{F}_\kappa^d)$ is no bigger than n, choose any signed set of points $Y = (Y^+, Y^-)$ such that $\text{card}(Y) > n$. Since the chambers of A^* form a partition of $\mathbf{R}^d$, then the maximum number of chambers possible is n (since A^* is in general position and $\text{card}(A^*) = \kappa$). There are more than n points, and regardless of the arrangement of points, at least one chamber will have more than one point in it. Therefore, for each set Y of greater than n points, one of the dichotomies will result in a $y^+ \in Y^+$ and a $y^- \in Y^-$ existing in the same chamber. In other words, A^* fails to shatter any set of $n+1$ points. So, $VC(\mathcal{F}_\kappa^d) < n+1$.

Combining yields $VC(\mathcal{F}_\kappa^d) = n = \text{card}(C(A^*))$. □

While it is true that a fixed ANN may define a set of hyperplanes that are not in general position (it is possible that there will be redundant processors or processors that yield parallel hyperplanes), the more meaningful situation is one where an ANN architecture is pushed to its capacity resulting in an arrangement of hyperplanes with maximal chambers. Hence, Theorem 14 describes a technique for evaluating the V-C dimension for a set of ANN architectures based purely on geometric properties of the hyperplanes. In the case where an arrangement is not in general position, the capacity of that ANN would be less than the V-C dimension of $\mathcal{F}_\kappa^d$.

5.4.4 Duality of Points and Hyperplanes. The reformulation of the V-C dimension falls out cleanly because a relationship exists naturally between the arrangement of the hyperplanes and the arbitrary (an unsigned) set of points in $\mathbf{R}^d$. For V-C dimension, the fact that only one set of points that is shatterable must exist in order to value the V-C dimension as that set's cardinality also contributes to the unique relationship. The proof essentially equates a chamber with a point, allowing

for the tie between cardinality of a set of chambers and the cardinality of the set of points.

It is not always the case that there exists a relationship between an invariant of the arrangement and the signed set. Other examples require a little finesse by appealing to the duality of hyperplanes and points. Zaslavsky uses this duality often to achieve many of the results of his counting arguments (50:4). Hence, where invariant analysis, thus far, has been about the structure of the arrangement of hyperplanes, by duality, it can also be about a set of points. This is accomplished through a mapping that equates hyperplanes in a $d-1$ dimensional subspace to points in a d dimensional vector space. This is an important tie because the new research presented in Chapter VI and Chapter VII is based on invariant analysis of signed sets of points.

5.4.5 The Lattice of the ANN Chamber Set. Recall that Theorem 14 establishes a relationship between the chambers and ANN capability analysis based on V-C dimension. Note that while a set of chambers is referred to as an invariant in the literature, it also possesses a lattice structure. The set of chambers produced by an ANN can be defined in this context.

Let $H^*(A) = \{h_1, h_2, h_3, ..., h_k, h_1^c, h_2^c, h_3^c, ..., h_k^c\}$, where h_i^c denotes the interior of the complement of h_i. Then, the set of chambers of an arrangement, A, is defined by

$$C(A) \equiv \left\{ \bigcap_{h^* \in T} h^* \ : \ \bigcap_{h^* \in T} h^* \neq \emptyset \text{ for all } T \subseteq H^*(A) \right\}.$$

Define $\preceq$ as set inclusion. That is, for $x, y \in C(A)$, $x \preceq y$ if and only if $x \subseteq y$. Then, the meet, $\wedge$, is defined as

$$x \wedge y = \text{glb}\{x, y\} = x \cap y.$$

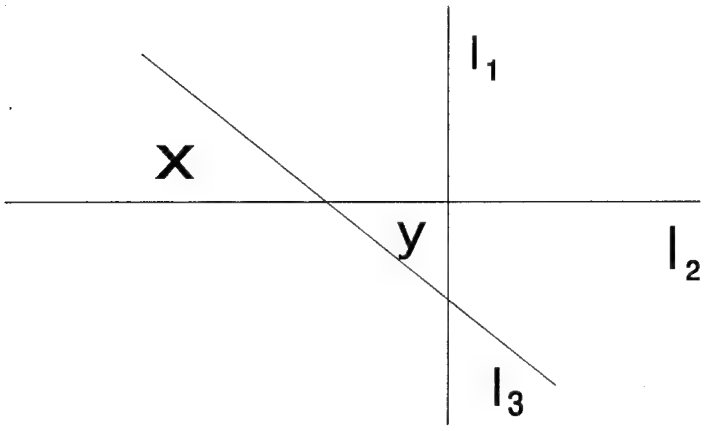


Figure 5. Chambers x and y.

By definition of $C(A)$, the meet operation is a closed operation on $C(A)$. The join, $\vee$, is defined as

$$x \vee y = \text{lub}\{x, y\} = x \cup y.$$

Note that the set $C(A)$ is not closed with respect to the join operation as demonstrated in Figure 5. In the figure, the set $x \cup y$ is not contained in $C(A)$. Note that the chamber $x = h_1^c \cap h_2 \cap h_3^c$, and the chamber $y = h_1^c \cap h_2^c \cap h_3$.

Theorem 15 *Let A be an arrangement of hyperplanes generated by an ANN. Let $\preceq$ be set inclusion and $\wedge$ and $\vee$ as defined above. Then, $(C(A), \preceq, \wedge, \vee)$ is a meet semi-lattice.*

Proof. First, show $(C(A), \preceq)$ is a partially ordered set. Let $x, y, z \in C(A)$. Since $x \subseteq x$, then $x \preceq x$. Hence, $\preceq$ is reflexive.

Assume $x \preceq y$ and $y \preceq x$. This implies $x \subseteq y$ and $y \subseteq x$, which implies $x = y$. Hence, $\preceq$ is antisymmetric.

Assume $x \preceq y$ and $y \preceq z$. This implies $x \subseteq y$ and $y \subseteq z$, which implies $x \subseteq z$. Hence $x \preceq z$, i.e. $\preceq$ is transitive.

Since $\preceq$ is reflexive, antisymmetric, and transitive for all $x, y, z \in C(A)$, then $(C(A), \preceq)$ is a partially ordered set. Moreover, $C(A)$ is, closed with respect to the meet operation, $\wedge$, which implies $(C(A), \preceq, \wedge, \vee)$ is a meet semi-lattice. □

Note that $(C(A), \preceq, \wedge, \vee)$ is not a lattice since the join is not closed. However, it is possible to produce a lattice on the chambers produced by the hyperplane arrangement of an ANN. Consider an architecture with an additional hidden-layer such that the interconnection weights are binary. This produces an additional *logic operation.* With the additional hidden-layer, the set of sets represented includes $C(A)$ and the union of all the elements of $C(A)$ which, by definition, is the power set of $C(A)$, denoted $\mathcal{P}(C(A))$.

Theorem 16 *Given an arrangement $A \in \mathbf{A}^d$, corresponding set, $C(A)$, ordering, $\preceq$, and meet and join operations as defined above, $(\mathcal{P}(C(A)), \preceq, \wedge, \vee)$ is a lattice.*

Proof. First, show $(\mathcal{P}(C(A)), \preceq)$ is a partially ordered set. Let $x, y, z \in \mathcal{P}(C(A))$. Since $x \subseteq x$, then $x \preceq x$. Hence, $\preceq$ is reflexive. Assume $x \preceq y$ and $y \preceq x$. This implies $x \subseteq y$ and $y \subseteq x$, which implies $x = y$. Hence, $\preceq$ is antisymmetric. Assume $x \preceq y$ and $y \preceq z$. This implies $x \subseteq y$ and $y \subseteq z$, which implies $x \subseteq z$. Hence $x \preceq z$, i.e. $\preceq$ is transitive.

Since $\preceq$ is reflexive, antisymmetric, and transitive for all $x, y, z \in \mathcal{P}(C(A))$, then $(\mathcal{P}(C(A)), \preceq)$ is a partially ordered set.

Moreover, by definition, $\mathcal{P}(C(A))$ is, closed with respect to the meet and the join operations, which implies $(\mathcal{P}(C(A)), \preceq, \wedge, \vee)$ is a lattice. □

5.5 Conclusions

In summary, this chapter described the required concepts of lattice theory and combinatorial geometry that facilitated an investigation of the lattice structures of feed-forward, single hidden-layer, perceptron artificial neural networks. This carved

the way for establishing that V-C dimension can be equated with the invariant chamber cardinality. Additionally, the cut-intersection semi-lattice of feed-forward, single hidden-layer, perceptron artificial neural networks was defined which facilitated the definition of the characteristic polynomial. Moreover, the set of chambers produced by an ANN was also found to be a semi-lattice.

VI. Generalized Invariant Analysis Applied to Artificial Neural Network Capability Analysis

This chapter pulls together the concepts of lattice theory and combinatorial geometry outlined in Chapter V and the desired properties of a quantifier of ANN capabilities outlined in Chapter IV. A framework will be built that generalizes methods used to determine ANN capabilities. The purpose of deriving the generalized framework is to build a method for comparing alternative architectures for their parsimonious solutions to the classification problem. In other words, a partial ordering of sets of sets derived from different architectures is sought. The ordering, to be presented, is determined by the *complexity* of signed sets for which an architecture can provide a solution. Hence, the crux of ANN capability analysis is reduced to characterizing this notion of complexity about signed sets. This is achieved through *invariant analysis.*

This view of capabilities analysis is consistent with V-C dimension. In other words, V-C dimension analysis can be posed as a specific instantiation of the generalized framework where cardinality is the invariant and V-C dimension is the function. However, the premise of this work is that V-C dimension has significant faults and, with the proper mathematics, more appropriate quantifiers can be designed. Specifically, the Ox-Cart dimension will be defined in Chapter VII which is based on an invariant called the geometric complexity.

Before any of the ANN capabilities analysis can be attempted, it is necessary to have a clear and concise definition of the problem space. Therefore, the definition and properties of the set of signed sets is investigated thoroughly.

6.1 Generalizing the Problem Space

In this section, the mathematical structure of the set of signed sets will be presented. In addition, the desired invariant properties about signed sets will be

defined. Moreover, these properties will be used to show the invariant nature of the mapping, geometric complexity.

6.1.1 The Collection of Signed Sets. The domain of the mappings that will define the set of invariants in the generalized framework is the set of signed sets. Recall that an ordered pair, (x, y), is a set of sets $\{\{x\}, \{x, y\}\}$ and $(x, y) \neq (y, x)$ unless $x = y$. Now, consider the following definition of a signed set.

Definition. A *signed set*, X^s, on $\mathbf{R}^d$, is an ordered pair of sets, $(X^+, X^-) \in \mathcal{P}(\mathbf{R}^d) \times \mathcal{P}(\mathbf{R}^d)$ such that $X^+ \cap X^- = \emptyset$. The corresponding *unsigned set* of X^s is $X = X^+ \cup X^-$.

Note that since $X^s = (X^+, X^-)$ is an ordered pair and $X^+ \cap X^- = \emptyset$, $(X^+, X^-) \neq (X^-, X^+)$ unless X^s is the *signed empty set* (which is just the empty set and will be denoted $\emptyset^s$). Hence, consider the rigorous definition of an ordered pair to establish *signed set equality,* denoted $\cong$. The ordered pair (X^+, X^-) can be uniquely defined by $\{\{X^+\}, \{X^+, X^-\}\}$.

Definition. Two signed sets, X_1^s and X_2^s, are said to be equal, written, $X_1^s \cong X_2^s$ if and only if $X_1^+ = X_2^+$ and $X_1^- = X_2^-$.

Let $\mathcal{X}$ denote the set of all signed sets on $\mathbf{R}^d$, i.e., $\mathcal{X} = \{X^s : X^s \text{ is a signed set}\}$. Define the zero signed element in $\mathcal{X}$ as $\emptyset^s = (\emptyset, \emptyset)$.

Consider the following definition of scalar multiplication on $\mathcal{X}$.

Definition. Signed set scalar multiplication is defined as

$$\alpha X^s \cong (\alpha X^+, \alpha X^-)$$

for all $\alpha \in \mathbf{R}$, $\alpha \neq 0$ and $X_1^s, X_2^s \in \mathcal{X}$. Recall that $\alpha X = \{\alpha x : x \in X\}$ for all $X \subset \mathbf{R}^d$.

Lemma 6 *The set $\mathcal{X}$ is closed with respect to scalar multiplication.*

Proof. Assume not. That is, without loss of generality assume $\alpha X^+ \cap \alpha X^- \neq \emptyset$ for some $\alpha \in \mathbf{R}$, $\alpha \neq 0$ and $X^s \in \mathcal{X}$. This implies that there exists $x_\alpha \in \alpha X^+$ and $x_\alpha \in \alpha X^-$. This implies there exists x such that $x_\alpha = \alpha x$ and, moreover, that $x \in X^+$ and $x \in X^-$. This implies $X^+ \cap X^- \neq \emptyset$ which is a contradiction. Hence, $\alpha X^+ \cap \alpha X^- = \emptyset$. Therefore, $\mathcal{X}$ is closed with respect to scalar multiplication. □

6.1.2 Invariant Properties for Operations on Signed Sets. The generalized framework for determining ANN capabilities will be centered on invariance. The following transformations, defined on $\mathcal{X}$, will help formalize mathematically the invariance desired of a mapping. Specifically, it is desired that the mappings that characterize signed sets will be invariant to dilation, translation, or rotation of the signed sets.

Let $\mathcal{Y}$ be the collection of finite sets Y such that $Y \subset \mathbf{R}^d$ that is

$$\mathcal{Y} = \{Y \subset \mathbf{R}^d : \text{card}(Y) \text{ is finite}\}.$$

Definition. Let $\gamma \in \mathbf{R}^+$. Define $D_\gamma : \mathcal{P}(\mathbf{R}^d) \to \mathcal{P}(\mathbf{R}^d)$ as

$$D_\gamma(Y) = \{\gamma y \in \mathbf{R}^d : y \in Y\}$$

for all $Y \in \mathcal{P}(\mathbf{R}^d)$. Then, the dilation operator, $\mathbf{D}_\gamma : \mathcal{X} \to \mathcal{X}$, is defined as

$$\mathbf{D}_\gamma(X^s) \doteq (D_\gamma(X^+), D_\gamma(X^-)),$$

for all $X^s \in \mathcal{X}$.

Definition. Let $x_0 \in \mathbf{R}^d$. Define $T_{x_0} : \mathcal{P}(\mathbf{R}^d) \to \mathcal{P}(\mathbf{R}^d)$ as

$$T_{x_0}(Y) = \{(x_0 + y) \in \mathbf{R}^d : y \in Y\}.$$

for all $Y \in \mathcal{P}(\mathbf{R}^d)$. Then, the translation operator, $\mathbf{T}_{x_0} : \mathcal{X} \to \mathcal{X}$, is defined as

$$\mathbf{T}_{x_0}(X^s) \doteq (T_{x_0}(X^+), T_{x_0}(X^-)),$$

for all $X^s \in \mathcal{X}$.

Definition. Let $\lambda \in \mathbf{R}^{d-1}$. Define $W_\lambda : \mathcal{P}(\mathbf{R}^d) \to \mathcal{P}(\mathbf{R}^d)$ as

$$W_\lambda(Y) = \{r_\lambda(y) \in \mathbf{R}^d : y \in Y\},$$

for all $Y \in \mathcal{P}(\mathbf{R}^d)$ where $r_\lambda : \mathbf{R}^d \to \mathbf{R}^d$ is a vector rotation operator that can be represented by an orthogonal matrix multiplication with the angle $\lambda = (\lambda_1, \lambda_2, ...\lambda_{d-1})$. Then, the rotation operator, $\mathbf{W}_\lambda : \mathcal{X} \to \mathcal{X}$, is defined as

$$\mathbf{W}_\lambda(X^s) \doteq (W_\lambda(X^+), W_\lambda(X^-)),$$

for all $X^s \in \mathcal{X}$.

Note that both $\mathbf{D}_\gamma$ and $\mathbf{W}_\lambda$ are linear. However, $\mathbf{T}_{x_0}$ is affine.

Convex hulls of sets will also be required.

Definition. Let $\alpha_1, \alpha_2, ..., \alpha_n \in \mathbf{R}^+$, such that

$$\sum_{i=1}^{n} \alpha_i = 1.$$

The convex hull of a set $Y \in \mathcal{Y}$, denoted $co(Y)$, is the set

$$co(Y) = \left\{ \sum_{i=1}^{n} \alpha_i y_i \in \mathbf{R}^d : \{y_1, y_2, ...y_n\} \subset Y,\ n \in \mathbf{N},\ \sum_{i=1}^{n} \alpha_i = 1,\ \alpha_i \geq 0 \right\}.$$

Let $\mathcal{X}^f$ denote the set of finite signed sets in $\mathcal{X}$.

Definition. Define the convex hull of a signed set $X^s = (X^+, X^-) \in \mathcal{X}^f$ to be $co(X^s) = (co(X^+), co(X^-))$.

It will be important for each of the desired invariant properties defined as transformations above to preserve convexity. Hence, consider the following lemmas.

Lemma 7 *Let $\gamma \in \mathbf{R}^+$. Then, for each $X^s \in \mathcal{X}^f$, $\mathbf{D}_\gamma(co(X^s)) \cong co(\mathbf{D}_\gamma(X^s))$.*

Proof. Let $\gamma \in \mathbf{R}^+$, $Y \in \mathcal{Y}$ with $card(Y) = n$. First, show $D_\gamma(Y)$ preserves convexity, i.e. $D_\gamma(co(Y)) = co(D_\gamma(Y))$.

$$\begin{aligned} D_\gamma(co(Y)) &= D_\gamma(\{\textstyle\sum_{i=1}^n \alpha_i y_i : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}) \\ &= \{\gamma \textstyle\sum_{i=1}^n \alpha_i y_i : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}\} \\ &= \{\textstyle\sum_{i=1}^n \alpha_i \gamma y_i : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}\} \\ &= co(D_\gamma(Y)). \end{aligned}$$

Therefore, for each $X^s \in \mathcal{X}^f$

$$\begin{aligned} \mathbf{D}_\gamma(co(X^s)) &\cong \mathbf{D}_\gamma(co(X^+), co(X^-)) \\ &\cong (D_\gamma(co(X^+)), D_\gamma(co(X^-))) \\ &\cong (co(D_\gamma(X^+), co(D_\gamma(X^-)) \\ &\cong co(\mathbf{D}_\gamma(X^s)). \qquad \square \end{aligned}$$

Lemma 8 *Let $x_0 \in \mathbf{R}^d$. Then, for each $X^s \in \mathcal{X}^f$, $\mathbf{T}_{x_0}(co(X^s)) \cong co(\mathbf{T}_{x_0}(X^s))$.*

Proof. Let $x_0 \in \mathbf{R}^d$, $Y \in \mathcal{Y}$ with $card(Y) = n$. First, show $T_{x_0}(Y)$ preserves convexity, i.e. $T_{x_0}(co(Y)) = co(T_{x_0}(Y))$. Consider

$$\begin{aligned} T_{x_0}(co(Y)) &= T_{x_0}(\{\textstyle\sum_{i=1}^n \alpha_i y_i : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}) \\ &= \{x_0 + \textstyle\sum_{i=1}^n \alpha_i y_i : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}\} \\ &= \{\textstyle\sum_{i=1}^n \alpha_i x_0 + \sum_{i=1}^n \alpha_i y_i : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}\} \\ &= \{\textstyle\sum_{i=1}^n \alpha_i (x_0 + y_i) : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}\} \\ &= co(T_{x_0}(Y)). \end{aligned}$$

Therefore, for each $X^s \in \mathcal{X}^f$

$$\begin{aligned}
\mathbf{T}_{x_0}(co(X^s)) &\cong \mathbf{T}_{x_0}(co(X^+), co(X^-)) \\
&\cong (T_{x_0}(co(X^+)), T_{x_0}(co(X^-))) \\
&\cong (co(T_{x_0}(X^+), co(T_{x_0}(X^-)) \\
&\cong co(\mathbf{T}_{x_0}(X^s)). \qquad \square
\end{aligned}$$

Lemma 9 *Let $\lambda \in \mathbf{R}^{d-1}$. Then, for each $X^s \in \mathcal{X}$, $\mathbf{W}_\lambda(co(X^s)) \cong co(\mathbf{W}_\lambda(X^s))$.*

Proof. Let $\lambda \in \mathbf{R}^{d-1}$, $Y \in \mathcal{Y}$ with $card(Y) = n$. First, show $W_\lambda(Y)$ preserves convexity, i.e. $W_\lambda(co(Y)) = co(W_\lambda(Y))$. Consider

$$\begin{aligned}
W_\lambda(co(Y)) &= W_\lambda(\{\textstyle\sum_{i=1}^n \alpha_i y_i : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}) \\
&= \{r_\lambda(\textstyle\sum_{i=1}^n \alpha_i y_i) : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}\} \\
&= \textstyle\sum_{i=1}^n \alpha_i r_\lambda(y_i) : y_i \in Y \ \sum_{i=1}^n \alpha_i = 1, \ \alpha_i \geq 0\}\} \\
&= co(W_\lambda(Y)).
\end{aligned}$$

Therefore, for each $X^s \in \mathcal{X}^f$

$$\begin{aligned}
\mathbf{W}_\lambda(co(X^s)) &\cong \mathbf{W}_\lambda(co(X^+), co(X^-)) \\
&\cong (W_\lambda(co(X^+)), W_\lambda(co(X^-))) \\
&\cong (co(W_\lambda(X^+), co(W_\lambda(X^-)) \\
&\cong co(\mathbf{W}_\lambda(X^s)). \qquad \square
\end{aligned}$$

This section has established a basic knowledge of signed sets which represent the classification problem.

6.2 Generalizing Artificial Neural Network Capability Quantifiers

Measurements of capability about ANNs can be viewed as functions of invariants about an arrangement A. The point is to broaden the understanding of how to characterize the strength of an arrangement which corresponds to an ANN ar-

chitecture. The goal is to identify mathematical entities that vary as the systems complexity varies and are invariant to characterizations of a system which do not contribute to an accurate description of an ANN's ability to solve classification problems.

Continuously throughout the literature (and in this dissertation), reference is made to the goal of *measuring* ANN capabilities. Also, V-C based functions are sometimes referred to as *measures.* Each of these terms is used loosely. In fact, V-C based quantifiers are not measures in the mathematical sense. Moreover, the quantifiers that will be defined by the general framework define a *function* of some invariant. As it turns out, the invariants are *semi-measures.* Consider the definition of a measure and a semi-measure.

Definition. Let $\mathcal{T}$ be a collection of sets Y such that $Y \subset \mathbf{R}^d$. A mapping, m, defined on a set $Y \in \mathcal{T}$ such that $m(Y) \in \mathbf{R}$ is a measure if it has the following properties:

1. $m(\emptyset) = 0$ (where $\emptyset$ is the empty set).
2. $m(Y) \geq 0$ for all nonempty $Y \in \mathcal{T}$.
3. For any finite set of finite disjoint sets, $\{Y_1, Y_2, ...Y_n\} \subset \mathcal{T}$,

$$m\left(\bigcup_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} m(Y_i).$$

Definition. A semi-measure is a mapping m defined on a set Y such that $m(Y) \in \mathbf{R}$ and has Properties 1 and 2 defined in above definition.

V-C dimension is not a measure since it fails to satisfy Property 3. In fact, this is a property that is undesirable for ANN capability quantifiers. As an example, consider two identical ANNs, each with only one perceptron that corresponds to the x-axis in $\mathbf{R}^2$. The addition of the two nets produces the same set of hyperplanes,

namely, the x-axis and the V-C dimension is the same. Hence the V-C dimension is not additive.

6.2.1 The Set of Invariants. In this section, the notion of invariants will be formalized. A set of invariants will be defined. Consider the following definition of an invariant. (Note that this definition is specific to this dissertation.)

Let $\mathcal{Z} = \{I : \mathcal{S} \to \mathbf{R} \mid I(\emptyset) = 0\}$ where $\mathcal{S}$ is a nonempty set. Then $\mathcal{Z}$ is a linear space over $\mathbf{R}$ with point-wise addition and scalar multiplication. Let $\mathcal{I} = \{I \in \mathcal{Z} \mid I(S) \geq 0 \text{ for all } S \in \mathcal{S}\}$.

Lemma 10 *$\mathcal{I}$ is a convex cone in $\mathcal{Z}$.*

Proof. Convexity: Let $\alpha \in [0,1]$ and $I_1, I_2 \in \mathcal{I}$, then by linearity of $\mathcal{Z}$,

$$[\alpha I_1 + (1-\alpha)I_2](S) = \alpha I_1(S) + (1-\alpha)I_2(S) \geq 0$$

for all $S \in \mathcal{S}$. Therefore $\alpha I_1 + (1-\alpha)I_2 \in \mathcal{I}$.

Cone: Let $\alpha \geq 0$, $I \in \mathcal{I}$, then $[\alpha I](S) = \alpha I(S) \geq 0$. So $\alpha I \in \mathcal{I}$. Hence, $\mathcal{I}$ is a convex cone in $\mathcal{Z}$. □

Definition. An element $I \in \mathcal{I}$, is said to be an *invariant with respect to a family*, $\mathcal{M}$, of mappings M, $M : \mathcal{I} \to \mathcal{I}$ if $M(I) = I$ for all $M \in \mathcal{M}$. That is

$$[M(I)](S) = I(S) \text{ for all } S \in \mathcal{S}.$$

Consider the following example. Recall the definition of the transformation operator $\mathbf{T}_{x_0}$.

Example. Let the invariant, $I \in \mathcal{I}$, be defined as, $I = \pi_1$, where $\pi_1 = \pi(L(A), 1)$, the characteristic polynomial evaluated at $t = 1$ defined on a cut-intersection semi-

lattice of an arrangement of hyperplanes, $L(A)$. Hence, $\mathcal{S} = L(A)$. Define $M \in \mathcal{M}$ as

$$M(\pi_1) = \pi_1 \circ T_{x_0}.$$

Define $\tilde{T}_{x_0}(L(A)) = \{T_{x_0}(a) : a \in L(A)\}$. The claim is that for some fixed $x_0 \in \mathbf{R}^d$, $M(\pi) = \pi$ for all $L(A)$. Given $L(A)$, consider

$$\begin{aligned}
[M(\pi_1)](L(A)) &= [\pi_1 \circ \tilde{T}_{x_0}](L(A)) \\
&= \pi_1(\tilde{T}_{x_0}(L(A))) \\
&= \pi(\tilde{T}_{x_0}(L(A), 1) \\
&= \text{card}(C(\tilde{T}_{x_0}(L(A)))) \\
&= \text{card}(C(L(A))) \\
&= \pi_1(L(A))
\end{aligned}$$

Hence, $M(\pi_1) = \pi_1$, for all $L(A)$. Therefore, π_1 is an invariant with respect to T_{x_0}. Note that this invariant maps arrangements to real numbers, i.e. $\pi_1 : L(A) \to \mathbf{R}$.

Consider another example of an invariant which does not rely on the existence of a cut-intersection semi-lattice or any other lattice. Instead it maps sets to real numbers.

Definition. Given $M : S \to S$, define $M^T : \mathcal{I} \to \mathcal{I}$ as the transpose of M written

$$M^T(I) = I \circ M \text{ for all } I \in \mathcal{I}.$$

Example. Let $\mathcal{S} = \mathcal{T}$, a collection of finite subsets of $\mathbf{R}^d$. Let the invariant, $I \in \mathcal{I}$, on an unsigned set, $X \in \mathcal{T}$ be defined as, $I = \text{card}(\cdot)$. Let

$$\mathcal{M} = \{W_\lambda^T : \lambda \in \mathbf{R}^{d-1}\} \cup \{D_\gamma^T : \gamma \in \mathbf{R}^+\} \cup \{T_{x_0}^T : x_0 \in \mathbf{R}^d\}.$$

The claim is that $M^T(\text{card}(\cdot)) = \text{card}(\cdot)$ for all $M^T \in \mathcal{M}$. Consider $M = T_{x_0}$

$$\begin{aligned}[M^T(\text{card}(\cdot))](X) &= [\text{card}(\cdot) \circ T_{x_0}](X) \\ &= \text{card}(T_{x_0}(X)) \\ &= \text{card}(X).\end{aligned}$$

Hence, $T^T_{x_0}(\text{card}(\cdot)) = \text{card}(\cdot)$ for all $x_0 \in \mathbf{R}^d$. Similarly, $W^T_\lambda(\text{card}(\cdot)) = \text{card}(\cdot)$ for all $\lambda \in \mathbf{R}^{d-1}$ and for all $\gamma \in \mathbf{R}^+$, $D^T_\gamma(\text{card}(\cdot)) = \text{card}(\cdot)$. Additionally, note that $\text{card}(\cdot)$ is a semi-measure. Therefore, cardinality is an invariant with respect to $\mathcal{M}$.

Because, an ANN capability quantifier should be about signed sets, what is sought is a set of invariants and a family of mappings that are defined on signed sets. Hence, consider the following definition for a set of invariants on signed sets, $\mathcal{I}^s$, which will be used to define the geometric complexity and the Ox-Cart dimension in Chapter VII.

Definition. Let $\mathcal{S} = \mathcal{X}$, the set of signed sets on $\mathbf{R}^d$. Given $X^s \in \mathcal{X}$, define the *invariant on signed sets*, $I \in \mathcal{I}^s$, as, $I : \mathcal{X} \to \mathbf{Z}^+$. Let $\mathcal{M}^s = \{\mathbf{W}^T_\lambda, \mathbf{D}^T_\gamma, \mathbf{T}^T_{x_0}\}$ where $\lambda \in \mathbf{R}^{d-1}$, $\gamma \in \mathbf{R}^+$, and $x_0 \in \mathbf{R}^d$. That is, for $M^T \in \mathcal{M}^s$

$$M^T(I) = I \circ M$$

for all $X^s \in \mathcal{X}$.

6.2.2 The Generalized Capability Quantifier. Now that the generalized set of invariants on unsigned sets, $\mathcal{I}$, and on signed sets, $\mathcal{I}^s$, has been defined, the generalized quantifier of ANN capabilities can be defined. However, first recall the definition of a family, $\mathcal{F}^d_\kappa$, of ANNs

$$\mathcal{F}^d_\kappa = \{A = \{l_1, l_2, l_3, ..., l_\kappa\} \subset \mathbf{A}^d\}.$$

where $\mathbf{A}^d$ is the set of all arrangements in d dimensions.

Definition. Given $d \in \mathbf{N}$. Let $\mathcal{J} \subset \mathcal{I}$ be a subset of invariants. The generalized quantifier of ANN capabilities with respect to $\mathcal{J}$, $\nu_{\mathcal{J}}$, is a mapping

$$\nu_{\mathcal{J}} : \mathcal{P}(\mathbf{A^d}) \rightarrow \mathbf{Z}.$$

Definition. Given $d \in \mathbf{N}$. Let $\mathcal{J}^s \subset \mathcal{I}^s$ be a subset of invariants defined on signed sets. The generalized quantifier of ANN capabilities with respect to $\mathcal{J}^s$, $\nu^s_{\mathcal{J}}$, is a mapping

$$\nu^s_{\mathcal{J}} : \mathcal{P}(\mathbf{A^d}) \rightarrow \mathbf{Z}.$$

Note that it also makes sense to require that ν and ν^s be semi-measures. In fact, since the invariants are semi-measures, one can expect that if $\mathcal{F}^d_\kappa$ is empty, then $\nu(\mathcal{F}^d_\kappa) = 0$. Additionally, it will be required that $\nu(\mathcal{F}^d_\kappa) \geq 0$ for all $\mathcal{F}^d_\kappa \subset \mathbf{A}^d$ for any d. (Where it is clear the subscript on ν will be dropped.)

Consider the reformulation of the V-C dimension of $\mathcal{F}^d_\kappa$. It has already been established that the V-C dimension of a set of hyperplane arrangements is the invariant $\text{card}(C(A))$, cardinality of the number of chambers of the arrangement $A^* \in \mathcal{F}^d_\kappa$, which can be analytically determined from the characteristic polynomial as the $\pi(A, 1)$. In the context of the generalization, $\text{card}(C(A)) \in \mathcal{I}$ and $\nu(\mathcal{F}^d_\kappa) = \max\{\text{card}(C(A)) : A \in \mathcal{F}^d_\kappa\} = VC(\mathcal{F}^d_\kappa)$. It is important to recall at this point that the objective of this dissertation is to characterize ν^s, since it will be shown that the analysis of capabilities is more appropriately defined about signed sets instead of arbitrary sets.

6.2.3 The Generalized Partial Ordering and Resulting Lattice of ANNs. Given the generalized capability quantifiers, ν and ν^s define the ordering, $\leq_\nu$, and $\leq_{\nu^s}$ on $\mathbf{A}^d$ as follows.

Definition. Given A_1, $A_2 \in \mathbf{A}^d$, $A_1 \leq_{\nu^s} A_2$ if and only if $\nu^s(A_1) \leq \nu^s(A_2)$. (Similarly, for $\leq_\nu$.)

Note that these are *not*, in general, partial orderings since they are not antisymmetric. However, it is possible to produce a partial ordering using equivalence classes of hyperplane arrangements.

Definition. Given $A \in \mathcal{F}_\kappa^d$, define the equivalence class of A, denoted $[A]$, as

$$[A] = \{B \in \mathcal{F}_\kappa^d : \nu^s(A) = \nu^s(B)\}.$$

Definition. Given $\mathcal{F}_\kappa^d$, define the collection of equivalence classes of $\mathcal{F}_\kappa^d$, denoted $\mathcal{E}(\mathcal{F}_\kappa^d)$, as

$$\mathcal{E}(\mathcal{F}_\kappa^d) = \left\{[A] : A \in \mathcal{F}_\kappa^d\right\}.$$

Now the ordering, $\leq_{\nu^s}$, (and $\leq_\nu$) can be defined on the collection of equivalence classes, $\mathcal{E}(\mathcal{F}_\kappa^d)$.

Definition. Given $\mathcal{E}(\mathcal{F}_\kappa^d)$, then $[A] \leq_{\nu^s} [B]$ if and only if $\nu^s([A]) \leq \nu^s([B])$. (Similarly, for $\leq_\nu$).

Lemma 11 *Given $d, \kappa \in \mathbf{N}$, $(\mathcal{E}(\mathcal{F}_\kappa^d), \leq_{\nu^s})$ is a partially ordered set.*

Proof. Let $d, \kappa \in \mathbf{N}$. Let $[A], [B], [C] \in \mathcal{E}(\mathcal{F}_\kappa^d)$. Note that $\nu^s([A]) = \nu^s([A])$ for all $[A] \in \mathcal{E}(\mathcal{F}_\kappa^d)$. Hence, $[A] \leq_{\nu^s} [A]$. In other words, $\leq_{\nu^s}$ is reflexive.

Assume $[A] \leq_{\nu^s} [B]$ and $[B] \leq_{\nu^s} [A]$. This implies $\nu^s([A]) \leq \nu^s([B])$ and $\nu^s([B]) \leq \nu^s([A])$ implying that $\nu^s([A]) = \nu^s([B])$. Hence, $[A] = [B]$. In other words, $\leq_{\nu^s}$ is antisymmetric.

Assume $[A] \leq_{\nu^s} [B]$ and $[B] \leq_{\nu^s} [C]$. This implies $\nu([A]) \leq \nu([B])$ and $\nu^s([B]) \leq \nu^s([C])$ implying that $\nu^s([A]) \leq \nu^s([C])$. Hence, $[A] \leq_{\nu^s} [C]$. In other words, $\leq_{\nu^s}$ is transitive.

Since $\leq_{\nu^s}$ is reflexive, antisymmetric, and transitive, then, $(\mathcal{E}(\mathcal{F}_\kappa^d), \leq_{\nu_s})$ is a partially ordered set. □

Note that $(\mathcal{E}(\mathcal{F}_\kappa^d), \leq_\nu)$ is also a partially ordered set.

Now, enough structure has been put into place to define the lattice of feed-forward, single hidden-layer, perceptron artificial neural networks. This is the crux of the generalization theory that allows analysis of ANN capabilities based on invariants. The lattice structure provides evidence that the generalized approach provides a well-structured, well-behaved environment in which to couch capability quantifiers. Consider the following definitions of meet and join on $\mathcal{E}(\mathcal{F}_\kappa^d)$. Let $[A], [B] \in \mathcal{E}(\mathcal{F}_\kappa^d)$. Then,

$$[A] \wedge [B] = \text{glb}\{[A], [B]\} = \begin{cases} [A] & \text{if} \quad \nu^s([A]) \leq \nu^s([B]) \\ [B] & \text{if} \quad \nu^s([B]) \leq \nu^s([A]) \end{cases}$$

and

$$[A] \vee [B] = \text{lub}\{[A], [B]\} = \begin{cases} [A] & \text{if} \quad \nu^s([B]) \leq \nu^s([A]) \\ [B] & \text{if} \quad \nu^s([A]) \leq \nu^s([B]) \end{cases}.$$

Theorem 17 *For each $\kappa, d \in \mathbf{N}$, $(\mathcal{E}(\mathcal{F}_\kappa^d), \leq_{\nu^s}, \wedge, \vee)$ is a lattice.*

Proof. Note that $(\mathcal{E}(\mathcal{F}_\kappa^d), \leq_{\nu^s})$ is a partially ordered set by Lemma 11. By construction of $\wedge$ and $\vee$, $\mathcal{E}(\mathcal{F}_\kappa^d)$ is closed with respect to $\wedge$ and $\vee$. Hence $(\mathcal{E}(\mathcal{F}_\kappa^d), \leq_{\nu^s}, \wedge, \vee)$ is a lattice for all $k, d \in \mathbf{N}$. □

Note that, again, a similar lattice can be defined based on analysis of unsigned sets. That is $(\mathcal{E}(\mathcal{F}_\kappa^d), \leq_\nu, \wedge, \vee)$ is also a lattice. However, to emphasize the point, the lattice based on signed sets is the one of interest in this dissertation.

6.3 Conclusions

In summary, what has been provided is a method for characterizing ANN capabilities that will allow comparisons of architectures based on their ability to implement the dichotomies of signed sets. This is formalized by the lattice of ANN's based on a generalized capability quantifier and partial ordering. To facilitate this, a generalized framework has been presented that characterizes the invariance of

signed sets. V-C dimension was posed as an instantiation of the generalization, and in Chapter VII is another instantiation: the Ox-Cart dimension.

VII. The Ox-Cart Dimension and the Lattice of Artificial Neural Networks

This chapter presents a means of determining the capability of feed-forward, single hidden-layer, perceptron, artificial neural networks to solve classification problems based on the complexity of a signed set, X^s. By example, the approach presented here demonstrates the usefulness of the generalized approach defined in Chapter VI. This notion of complexity is encased in a mapping called the *geometric complexity,* denoted GC. The capability quantifier is called the Ox-Cart dimension, denoted $OC(A)$. It is defined on an arrangement, A, and on a set of arrangements of hyperplanes generated from an ANN. A partial ordering of ANNs will be defined using the Ox-Cart dimension. The ordering provides a comparison of ANNs based on their ability to solve classification problems determined by the complexity of the problem. Moreover, that ordering will result in a lattice defined on ANNs.

Since the ordering is determined by the *complexity* of signed sets for which an architecture can provide a solution, the crux of ANN capability analysis is reduced to characterizing this notion of complexity about signed sets. This is achieved through *invariant analysis.* In particular, GC will be posed as an invariant in the generalized framework defined in Chapter VI and OC as the function which operates on that invariant. Hence, GC is a mapping on signed sets and OC is a mapping on arrangements.

It should be noted that defining the Ox-Cart dimension within the generalized framework directs the capabilities analysis at specific geometric characterizations of classification problems, not the cardinality of arbitrary classification problems. The specificity is determined by the geometric complexity mapping GC. This distinction is the major difference in the Ox-Cart dimension and the V-C dimension.

The geometric complexity mapping is centered around the notion that the iterative intersections of the signed sets' convex hulls are an appropriate indicator of

the difficulty of the classification problem. It should be emphasized that the Ox-Cart dimension ultimately appeals to the analysis of a particular problem. This differs from V-C dimension based methods which are based on arbitrary arrangements of sets (not signed sets). By moving the analysis to signed sets from unsigned sets, the methods become more useful in applications, since they provides a tool for directly analyzing an ANN's requirements for solving a classification problem. Moreover, the sophistication of algorithms for deriving the geometric invariants used to determine the Ox-Cart dimension makes evaluating the function accessible. (36)

The structure of this chapter parallels that of Chapter VI. Where Chapter VI was general, this chapter will provide an instantiation. First, there will be a discussion of geometric complexity as an invariant and its properties will be investigated. Then, the Ox-Cart dimension will be defined as a particular ANN capability quantifier, ν. A partial ordering about the Ox-Cart dimension will be defined and the lattice that it produces will be established. An example is also presented. Finally, a comparison of the V-C dimension and the Ox-Cart dimension is presented.

7.1 The Geometric Complexity Mapping

This section will provide a definition of geometric complexity, examples, the discontinuity of GC, and its invariant properties.

7.1.1 Definition of GC. The geometric complexity is a mapping from the set of finite signed sets, $\mathcal{X}^f$, to a nonnegative integer. That is,

$$GC : \mathcal{X}^f \rightarrow \mathbf{Z}^+.$$

The value of $GC(X^s)$ is indicative of how "mixed up" the dichotomy of X^s is, which should correlate with the difficulty of separating the two sets, X^+, X^- with hyperplanes (difficulty being defined as the number of hyperplanes required to separate the sets). For the purpose of defining $GC(X^s)$, "being mixed up" is mathematically

described by evaluating the intersections of the convex hulls of X^+ and X^-, denoted $co(X^+)$ and $co(X^-)$, respectively. This is done iteratively, breaking the classification problem, i.e. the signed set, down starting on the outside and moving inward. The higher the value of GC, the more difficult the problem.

The mapping, GC, is constructed from three mappings, S, h, and H.

Define the mapping $S : \mathcal{X}^f \rightarrow \{0, 1, 2\}$ as

$$S(X^s) = \begin{cases} 0 & \text{if} \quad X^s = \emptyset \\ 1 & \text{if} \quad X^+ = \emptyset \text{ xor } X^- = \emptyset \\ 2 & \text{if} \quad X^+ \neq \emptyset \text{ and } X^- \neq \emptyset. \end{cases}$$

$S(X^s)$ mathematically clarifies if one or both of the sets, X^+, X^- are empty. If $S(X^s) = 0$, then there is no hyperplane required to separate the empty set. If $S(X^s) = 1$, then no hyperplane is required to implement the dichotomy on X^s. However, if $S(X^s) = 2$, more analysis is required to determine complexity. So, form the convex hulls of X^+ and X^- that is $co(X^+)$ and $co(X^-)$. If $co(X^+) \cap co(X^-) = \emptyset$, then, by the Hahn-Banach Theorem, the set is separable by one hyperplane (31). If $co(X^+) \cap co(X^-) \neq \emptyset$, then analysis of the convex hulls must continue in order to determine the complexity within $co(X^+) \cap co(X^-)$. To express this mathematically, a binary mapping $h : \mathcal{X}^f \rightarrow \{0, 1\}$ is required. For $X^s \in \mathcal{X}^f$, define

$$h(X^s) = \begin{cases} 0 & \text{if} \quad co(X^+) \cap co(X^-) = \emptyset \\ 1 & \text{if} \quad co(X^+) \cap co(X^-) \neq \emptyset \end{cases}$$

To reduce X^s to the portion that has not been analyzed define $H : \mathcal{X}^f \rightarrow \mathcal{X}^f$ as

$$H(X^s) \cong (X^+ \cap [co(X^+) \cap co(X^-)], \; X^- \cap [co(X^+) \cap co(X^-)]).$$

If $X^s \cong (\emptyset, \emptyset)$, then, $H(X^s) \cong H(\emptyset, \emptyset) \cong (\emptyset, \emptyset)$. Also, since $H : \mathcal{X}^f \rightarrow \mathcal{X}^f$, the composition, $H \circ H$, is well-defined. For clarity, let H^k denote H composed k times.

Theorem 18 *Let $X^s \in \mathcal{X}^f$, then there exists a $k \in \mathbf{N}$ such that $H^k(X^s) \cong (\emptyset, \emptyset)$.*

Proof. This proof will be by contradiction. That is, for $X^s \in \mathcal{X}^f$, assume $H^k(X^s) \neq (\emptyset, \emptyset)$ for all $k \in \mathbf{N}$. Since X^s is finite, this implies that $H^k(X^s) \cong H^{k-1}(X^s)$ for some $k = k_0$. Let $Z \cong H^{k_0-1}(X^s)$. Then, $H(Z) \cong Z$. This implies

$$Z^+ \cap [co(Z^+) \cap co(Z^-)] \;=\; Z^+ \;\text{ and }\; Z^- \cap [co(Z^+) \cap co(Z^-)] \;=\; Z^-.$$

Hence,

$$co(Z^+) \cap co(Z^-) \;\supseteq\; Z^+ \;\text{ and }\; co(Z^+) \cap co(Z^-) \;\supseteq\; Z^-.$$

So that

$$co(co(Z^+) \cap co(Z^-)) \;\supseteq\; co(Z^+) \;\text{ and }\; co(co(Z^+) \cap co(Z^-)) \;\supseteq\; co(Z^-).$$

Therefore

$$co(Z^+) \cap co(Z^-) \;\supseteq\; co(Z^+) \;\text{ and }\; co(Z^+) \cap co(Z^-) \;\supseteq\; co(Z^-),$$

which implies $co(Z^+) = co(Z^-)$, implying that $Z^+ \cap Z^- \neq \emptyset$. This is a contradiction. Hence, there must exist $k^* \in \mathbf{N}$ such that $H^{k^*}(X^s) \cong (\emptyset, \emptyset)$. □

$H(X^s)$ reduces the domain of the analysis to the portion of the problem that has not been characterized for complexity. In this manner, the problem would be picked apart and values of $S(X^s)$ and $h(\dot{X}^s)$ would be accumulated until X^s is reduced to $(\emptyset, \emptyset)$. This process is housed in the function for $GC(X^s)$.

Finally, the *geometric complexity* of a finite signed set X^s can be defined.

Definition. The *geometric complexity* of a finite signed set X^s, $GC : \mathcal{X}^f \to \mathbf{Z}^+$, is defined as

$$\begin{aligned} GC(X^s) &= [S(X^s) + h(X^s)] + [S(H(X^s)) + h(H(X^s))] + \\ &\quad [S(H(H(X^s))) + h(H(H(X^s)))] + \ldots \\ &= \sum_{k=0}^{\infty} \left[S(H^k(X^s)) + h(H^k(X^s)) \right]. \end{aligned} \tag{17}$$

Theorem 19 *For each $X^s \in \mathcal{X}^f$, $GC(X^s) < \infty$. Hence, GC is defined on all of $\mathcal{X}^f$.*

Proof. By Theorem 18, $GC(X^s)$ is a finite sum of integers and is, therefore, finite for all $X^s \in \mathcal{X}^f$. □

7.1.2 Examples of GC. Consider the following examples which demonstrate the sensitivity of GC to different geometric situations. The first set of examples demonstrates that $GC(X^s)$ increases, as does the required number of hyperplanes to implement the dichotomy *irrespective of the cardinality of X^s.*

Example 1. For $X^s \cong \emptyset^s$,

$$\begin{aligned} GC(X^s) &= S(X^s) + h(X^s) \\ &= 0 + 0 \\ &= 0, \end{aligned}$$

since $H(\emptyset, \emptyset) \cong (\emptyset, \emptyset)$.

Example 2. For $X^s \cong (X^+, \emptyset)$, or $X^s \cong (\emptyset, X^-)$,

$$\begin{aligned} GC(X^s) &= S(X^s) + h(X^s) \\ &= 1 + 0 \\ &= 1, \end{aligned}$$

again since $H(\emptyset, \emptyset) \cong (\emptyset, \emptyset)$.

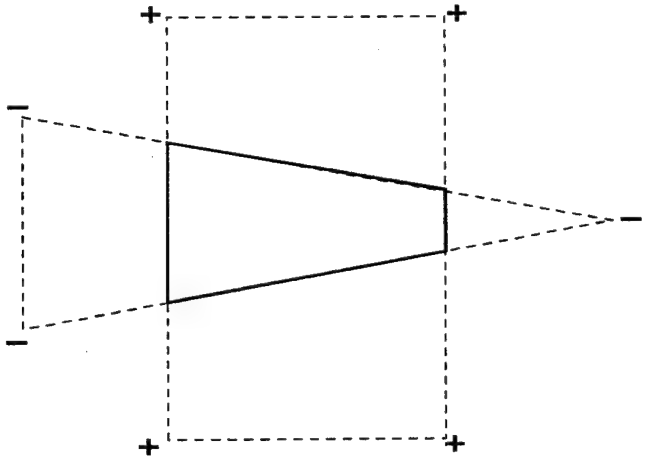


Figure 6. Example 4, GC(X)=3.

Example 3. For $X^s \cong (X^+, X^-)$, where $X^+ \neq \emptyset$ and $X^- \neq \emptyset$, but $co(X^+) \cap co(X^-) = \emptyset$, then

$$\begin{aligned} GC(X^s) &= S(X^s) + h(X^s) \\ &= 2 + 0 \\ &= 2, \end{aligned}$$

and again $H(\emptyset, \emptyset) \cong (\emptyset, \emptyset)$.

These three examples demonstrate in a very simple way how the geometric complexity of sets is determined.

The following examples will help explore the more interesting aspects of the mapping. Consider the sets and GC evaluation pictured in Figures 6 thru 9.

Example 4. For X^s, as shown in Figure 6,

$$\begin{aligned} GC(X^s) &= S(X^s) + h(X^s) \\ &= 2 + 1 \\ &= 3. \end{aligned}$$

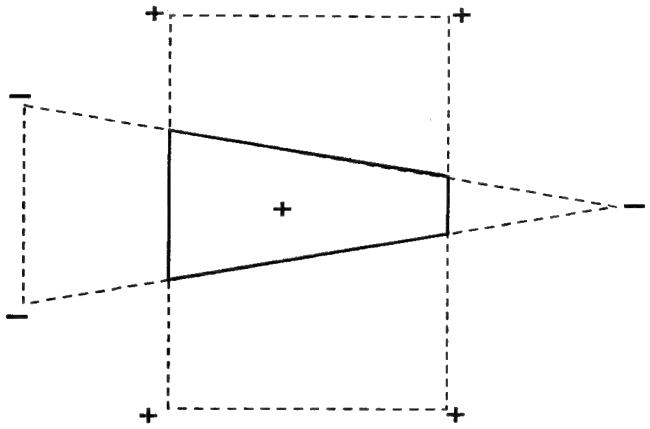


Figure 7. Example 5, GC(X)=4.

Example 5. For X^s, as shown in Figure 7,

$$\begin{aligned} GC(X^s) &= S(X^s) + h(X^s) + S(H(X^s)) + h(H(X^s)) \\ &= 2+1+1+0 \\ &= 4. \end{aligned}$$

Example 6. For X^s, as shown in Figure 8,

$$\begin{aligned} GC(X^s) &= S(X^s) + h(X^s) + S(H(X^s)) + h(H(X^s)) \\ &= 2+1+2+0 \\ &= 5. \end{aligned}$$

Example 7. For X^s, as shown in Figure 9,

$$\begin{aligned} GC(X^s) &= \textstyle\sum_{k=0}^{3} \left[S(H^k(X^s)) + h(H^k(X^s))\right] \\ &= 2+1+2+1+2+1 \\ &= 9. \end{aligned}$$

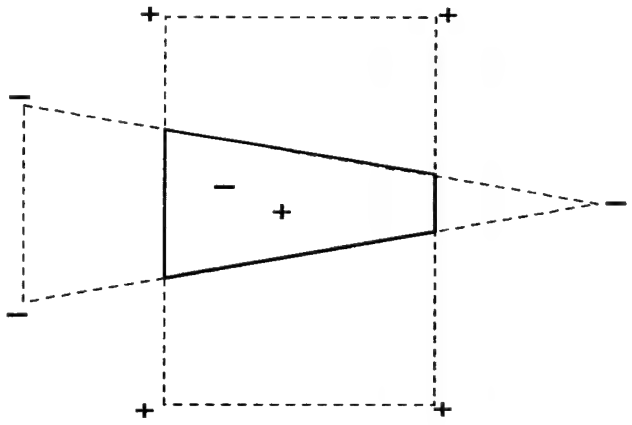


Figure 8. Example 6, GC(X)=5.

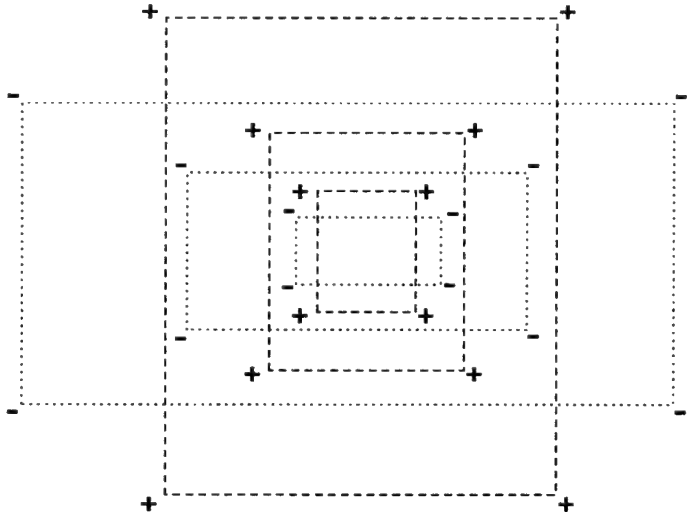


Figure 9. Example 7, GC(X)=9.

These examples demonstrate sets of increasing difficulty for ANN implementation that is represented in the increased values of GC. Once again, note that, in these examples, V-C dimension based quantifiers would grow disproportionately to the number of hyperplanes required to separate the points. This is simply because V-C analysis would be completed on the 2^n colored sets. Moreover, since O-C will be defined on signed sets, a comparison of the values of V-C dimension and the Ox-Cart dimension is meaningless. However, a comparison of the mappings of V-C dimension and the Ox-Cart dimension will be given in the chapter summary. This really just establishes that, in order to get an accurate picture of what is required of an ANN to solve a problem, V-C based quantifiers may be ambiguous. An approach, such as GC, may be more indicative of the true hyperplane arrangement requirements to implement the dichotomy of a signed set.

From these examples, it can be seen that, in some aspects, GC is cardinality insensitive. For instance, the value of GC for Examples 1 and 2 would not change if the cardinality of X^s changed. This is also true in the other examples unless the set X^s was changed so that there were an increased number of convex hull decomposition iterations, then GC would also change. Also, note that the function GC can theoretically be applied regardless of dimension. However, dimension affects the algorithms used to compute convex hulls and intersections.

7.1.3 Further Analysis of GC. It would seem intuitive that another mapping similar to GC could be derived mimicking Sontag's use of the Hausdorff metric to introduce continuity. However, continuity with respect to the Hausdorff metric represents a function's sensitivity to changes in the distance between the points in the signed set. This is inappropriate for GC since it is valued on geometrical concepts of signed sets. That is, a finite signed set's geometric complexity is not inherently dependent on the points' distances between each other.

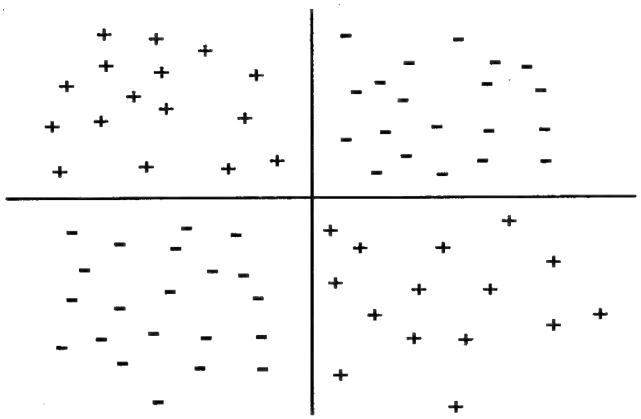


Figure 10. Example 8, the XOR problem.

Now, consider three examples (illustrated in Figures 10, 11 and 12) that demonstrate the sensitivity of GC with respect to changes in the geometry of the signed sets. GC gives ambiguous results for Examples 8 and 9.

Example 8. The classic *exclusive or* (XOR) problem (Figure 10) demonstrates that as the cardinality of the set increases, GC will also increase. However, only two hyperplanes will be required to separate the points.

Example 9. For Baum's example of the alternating n-gon problem (Figure 11), regardless of the cardinality of the set, GC will always be 2. However, the required number of lines increases as the cardinality of the set increases (see Theorem 10).

Example 10. The rings problem (Figure 12) shows that cardinality can be increased and GC will remain constant. This is reflective of the fact that the number of hyperplanes required will not grow with cardinality.

It would appear that changing GC to include a count of the vertices of the convex hulls (the cardinality of the extremal sets) could make the value of GC more in line with the solution to the n-gon problem. However, the GC value for the rings example would be inappropriate. In conclusion, in order to build a GC mapping so that it gives unambiguous results for all geometric situations, the algorithm would in essence have to determine exactly how many lines are required and where to place

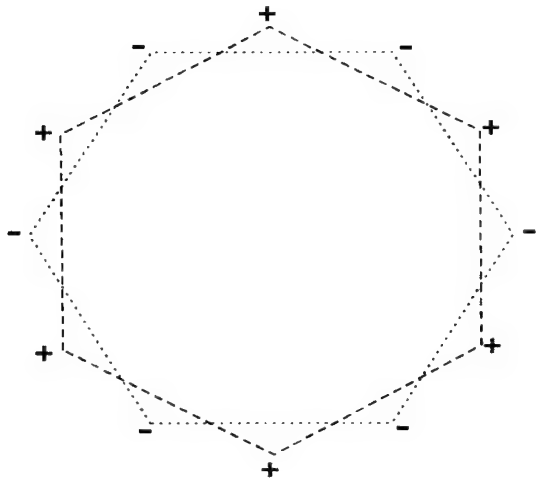


Figure 11. Example 9, the n-gon problem.

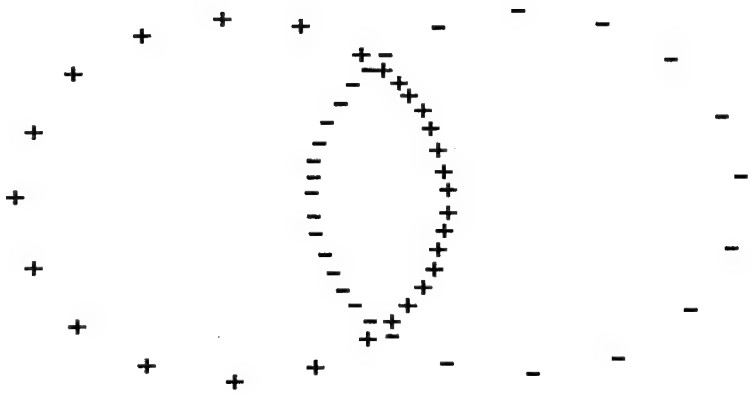


Figure 12. Example 10, the rings problem.

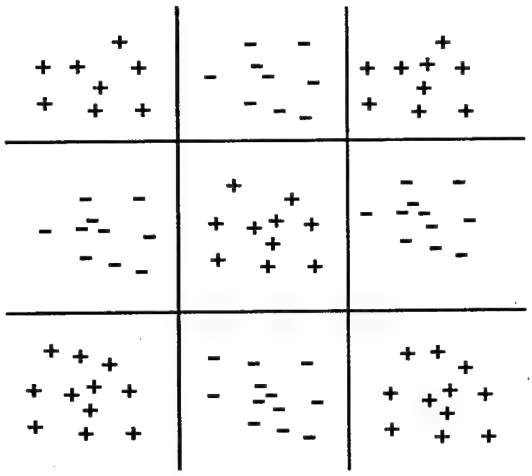


Figure 13. Example 11, the checkerboard problem.

them, which is the job of the ANN in the first place. So, in essence, it would be defeating the whole purpose not to mention being computationally unreasonable.

Consider one additional geometric situation, the checkerboard example.

Example 11. For the checkerboard problem, as shown in Figure 13, it is evident once again how it is possible for GC to grow unreasonably as the cardinality of the set grows.

7.1.4 GC is an Invariant. Mathematically, GC is a mapping from a signed set $X^s \cong (X^+, X^-)$, to a nonnegative integer $z \in \mathbf{Z}^+$,

$$GC : \mathcal{X}^f \rightarrow \mathbf{Z}^+.$$

The claim is that $GC \in \mathcal{I}^s$ for any $A \in \mathbf{A}^d$, for any d. To proves this, GC must be shown to be a semi-measure and be invariant with respect to $\mathbf{D}_\gamma$, $\mathbf{T}_{x_0}$, and $\mathbf{W}_\lambda$.

Consider the following lemma which establishes that GC is a semi-measure.

Lemma 12 *GC is a semi-measure, not a measure.*

Proof. By definition, $GC(\emptyset^s) = 0$ and $GC(X^s) \geq 0$ for all $X^s \in \mathcal{X}^f$. Hence, GC is a semi-measure.

However,

$$GC(\cup X^s) \neq \sum GC(X^s)$$

for a finite number of signed sets $X^s \in \mathcal{X}^f$. Consider two signed sets X_1^s and X_2^s in $\mathbf{R}^2$ such that both can be separated by the x-axes. Moreover, let X_1^s and X_2^s be such that X_1^+ and X_2^+ lie above the x-axes and X_1^- and X_2^- lie below the x-axes. Then, $GC(X_1^s) = GC(X_2^s) = 2$. However, $GC(X_1^s \cup X_2^s) = 2$ $(\neq 4)$. □

GC is advertised as an invariant. What exactly is invariant about it? The hope was that only changes in the complexity of the geometry of a signed set would affect a change in the value of GC. Hence, it would be expected that the mapping would be invariant to translational, rotational, and scale variations to X^s. Consider the following lemmas.

Lemma 13 *$GC(X^s) = GC(\mathbf{T}_{x_0}(X^s))$ for all $X^s \in \mathcal{X}^f$ and $x_0 \in \mathbf{R}^d$.*

Proof. Since GC is simply a sum of integers dictated by the mappings $S(X^s)$ and $h(X^s)$, then the proof is reduced to showing that these mappings are invariant for any signed set. Note that the results of H actually do change. However, it is the effect on S and h that matter to the value of GC.

Hence, consider $S(\mathbf{T}_{x_0}(X^s))$. If $X^s \cong \emptyset$, then $S(X^s) = 0$. Clearly, $\mathbf{T}_{x_0}(\emptyset) = \emptyset$ and $S(\mathbf{T}_{x_0}(\emptyset)) = 0$.

Similarly, for $X^s \in \mathcal{X}^f$ such that $S(X^s) = 1$, then $X^s \cong (X^+, \emptyset)$ or $X^s \cong (\emptyset, X^-)$. Then, $S(\mathbf{T}_{x_0}(X^s)) = S(T_{x_0}(X^+), \emptyset) = 1$ or $S(\mathbf{T}_{x_0}(X^s)) = S(\emptyset, T_{x_0}(X^-)) = 1$.

Finally, for $X^s \in \mathcal{X}^f$ such that $S(X^s) = 2$, then $co(X^+) \cap co(X^-) \neq \emptyset$. Therefore, $co(T_{x_0}(X^+)) \cap co(T_{x_0}(X^-)) \neq \emptyset$. Then, $S(\mathbf{T}_{x_0}(X^s)) = S(T_{x_0}(X^+), T_{x_0}(X^-)) = 2$. Hence, $S(X^s) = S(\mathbf{T}_{x_0}(X^s))$.

Now consider $h(\mathbf{T}_{x_0}(X^s))$. Then, $h(\mathbf{T}_{x_0}(X^s)) = h(T_{x_0}(X^+), T_{x_0}(X^-))$. Notice that

$$\begin{aligned} co(T_{x_0}(X^+)) \cap co(T_{x_0}(X^-)) &= T_{x_0}(co(X^+)) \cap T_{x_0}(co(X^-)) \\ &= T_{x_0}(co(X^+) \cap co(X^-)) \end{aligned}$$

by Lemma 8. If $co(X^+) \cap co(X^-) = \emptyset$, then $T_{x_0}(co(X^+) \cap co(X^-)) = \emptyset$, and if $co(X^+) \cap co(X^-) \neq \emptyset$, then $T_{x_0}(co(X^+) \cap co(X^-)) \neq \emptyset$. Hence, $h(X^s) = h(\mathbf{T}_{x_0}(X^s))$.

By combining, $GC(X^s) = GC(\mathbf{T}_{x_0}(X^s))$. □

Lemma 14 *$GC(X^s) = GC(\mathbf{D}_\gamma(X^s))$ for all $X^s \in \mathcal{X}^f$ and $\gamma \in \mathbf{R}^+$.*

Proof. The proof follows similar to Lemma 13 relying on Lemma 7. □

Lemma 15 *$GC(X^s) = GC(\mathbf{W}_\lambda(X^s))$ for all $X^s \in \mathcal{X}^f$ and $\lambda \in \mathbf{R}^{d-1}$.*

Proof. The proof follows similarly as Lemma 13 relying on Lemma 9. □

Enough material has been given to claim that GC is truly an invariant. Consider the following theorem.

Theorem 20 *$GC \in \mathcal{I}^s$ with respect to the set of mappings $\mathcal{M}^s$.*

Proof. By Lemmas 13, 14, and 15, $GC \in \mathcal{I}^s$. That is GC is an invariant. □

7.2 Definition of the Ox-Cart Dimension

In this section, the Ox-Cart dimension will be defined along with the ordering induced by the Ox-Cart dimension. It will be shown that each of these are instantiations of the general framework. That is the Ox-Cart dimension is an instantiation of ν^s and has the corresponding ordering facilitating the construction of a lattice.

Hence, through the Ox-Cart dimension, ANN architectures can be compared or ordered based on their ability to handle different levels of geometric complexity.

In accordance to the general framework defined in Chapter VI, the Ox-Cart dimension will be defined as a function of the invariant GC. Let $A \in \mathbf{A}^d$ for some $d \in \mathbf{Z}^+$. Then, the Ox-Cart dimension, OC, is a mapping from one specific arrangement to a positive integer. That is,

$$OC : \mathbf{A}^d \rightarrow \mathbf{Z}^+.$$

Let $\mathcal{X}_A^s \subset \mathcal{X}^s$ be the collection of signed sets which can be dichotomized by A.
Definition. The OC dimension of A is defined by

$$OC(A) = \sup\{GC(X^s) : X^s \in \mathcal{X}_A^f\}. \tag{18}$$

This is a quantifier on *an* arrangement of hyperplanes generated by an ANN. Define the ordering, $\preceq_{OC}$, of arrangements as follows. Let A_1, $A_2 \in \mathbf{A}^d$. Then, $A_1 \preceq_{OC} A_2$ if and only if $OC(A_1) \leq OC(A_2)$. Note that this is not antisymmetric. Hence, $\preceq_{OC}$ is not a partial ordering. However, in the following section a partial ordering is defined using equivalence classes.

7.3 The Lattice of Feed-forward, Single Hidden-Layer, Perceptron Artificial Neural Networks based on the Ox-Cart Dimension

Recall the definition of $\mathcal{F}_\kappa^d$,

$$\mathcal{F}_\kappa^d = \{A = \{l_1, l_2, ..., l_k\} \in \mathbf{A}^d : \ k = \kappa\}.$$

Consider the following definitions of equivalence classes defined specifically about the Ox-Cart dimension. (Boldface is used to annotate that the Ox-Cart dimension is being defined on a set of arrangements.)

Definition. Given $A \in \mathcal{F}_\kappa^d$, define the *equivalence class of* A, denoted $[A]$, as

$$[A] = \{B \in \mathcal{F}_\kappa^d : OC(A) = OC(B)\}.$$

Definition. Given $\mathcal{F}_\kappa^d$, define the *collection of equivalence classes of* $\mathcal{F}_\kappa^d$, denoted $\mathcal{E}(\mathcal{F}_\kappa^d)$, as

$$\mathcal{E}(\mathcal{F}_\kappa^d) = \{[A] : A \in \mathcal{F}_\kappa^d\}.$$

Let $\mathcal{X}_{\mathcal{F}_\kappa^d}^f \subset \mathcal{X}^f$ be the set of signed sets which can be implemented by some $A \in \mathcal{F}_\kappa^d$. Then, the O-C dimension on the collection of arrangements, $\mathcal{F}_\kappa^d$, can be defined by

$$\mathbf{OC}(\mathcal{F}_\kappa^d) = \sup\{GC(X^s) : X^s \in \mathcal{X}_A^f \text{ for all } A \in \mathcal{F}_\kappa^d\}. \tag{19}$$

The ordering, $\preceq_{\mathbf{OC}}$, can be defined on the collection of equivalence classes of hyperplanes, $\mathcal{E}(\mathcal{F}_\kappa^d)$ as follows. Let $[A]$, $[B] \in \mathcal{E}(\mathcal{F}_\kappa^d)$. Then, $[A] \preceq_{\mathbf{OC}} [B]$ if and only if $\mathbf{OC}([A]) \leq \mathbf{OC}([B])$.

Lemma 16 *$(\mathcal{E}(\mathcal{F}_\kappa^d), \preceq_{\mathbf{OC}})$ is a partially ordered set for all $k, d \in \mathbf{N}$.*

Proof. By Lemma 11. □

Structure has been put into place to define the lattice of feed-forward, single hidden-layer, perceptron artificial neural networks based on the Ox-Cart dimension. Consider the following definitions of meet and join on $\mathcal{E}(\mathcal{F}_\kappa^d)$. Let $[A], [B] \in \mathcal{E}(\mathcal{F}_\kappa^d)$. Then,

$$[A] \wedge [B] = glb\{[A], [B]\} = \begin{cases} [A] & \text{if} \quad \mathbf{OC}([A]) \leq \mathbf{OC}([B]) \\ [B] & \text{if} \quad \mathbf{OC}([B]) \leq \mathbf{OC}([A]) \end{cases}$$

and

$$[A] \vee [B] = lub\{[A], [B]\} = \begin{cases} [A] & \text{if} \quad \mathbf{OC}([B]) \leq \mathbf{OC}([A]) \\ [B] & \text{if} \quad \mathbf{OC}([A]) \leq \mathbf{OC}([B]) \end{cases}.$$

Theorem 21 *For each $d, \kappa \in \mathbf{N}$, then $(\mathcal{E}(\mathcal{F}_\kappa^d), \preceq_{\mathbf{OC}}, \wedge, \vee)$ is a lattice.*

Proof. Let $d, \kappa \in \mathbf{N}$, then $(\mathcal{E}(\mathcal{F}_\kappa^d), \preceq_{\mathbf{OC}})$ is a partially ordered set by Lemma 16. By the definition of $\wedge$ and $\vee$, $\mathcal{E}(\mathcal{F}_\kappa^d)$ is closed with respect to $\wedge$ and $\vee$. Hence, $(\mathcal{E}(\mathcal{F}_\kappa^d), \preceq_{\mathbf{OC}}, \wedge, \vee)$ is a lattice. □

Consider the following theorem.

Theorem 22 *If* $\mathbf{f}, \mathbf{g} \subset \mathbf{A}^d$ *such that* $\mathbf{f} \subseteq \mathbf{g}$, *then* $\mathbf{f} \preceq_{\mathbf{OC}} \mathbf{g}$ *for all* $d \in \mathbf{N}$.

Proof. Let $\mathbf{f}, \mathbf{g} \subset \mathbf{A}^d$. Assume $\mathbf{f} \subseteq \mathbf{g}$. Note that $\mathbf{f} \subseteq \mathbf{g}$ implies that $\mathcal{X}_\mathbf{f}^f \subseteq \mathcal{X}_\mathbf{g}^f$. Hence

$$\sup\{GC(X^s) : X^s \in \mathcal{X}_A^f, \ \forall\, A \in \mathbf{f}\} \leq \sup\{GC(X^s) : X^s \in \mathcal{X}_A^f, \ \forall\, A \in \mathbf{g}\}.$$

Therefore, $\mathbf{OC}(\mathbf{f}) \leq \mathbf{OC}(\mathbf{g})$. In other words, $\mathbf{f} \preceq_{\mathbf{OC}} \mathbf{g}$ for all $d \in \mathbf{N}$. □

In summary, what has been provided is a method for characterizing ANN capabilities that will allow comparisons of architectures based on the geometric complexity of signed sets.

7.3.1 An Example Using $\preceq_{\mathbf{OC}}$. Consider the following example which analyzes the difference in capability of ANNs that have nonzero valued thresholds at each node in the hidden-layer and those that have zero valued thresholds.

Consider $\mathcal{F}_\kappa^2$. Define the subsets $\mathbf{f}_\kappa, \mathbf{g}_\kappa \subset \mathcal{F}_\kappa^2$ by

$$\mathbf{f}_\kappa = \left\{ A \in \mathcal{F}_\kappa^2 : A = \bigcup_{i=1}^{\kappa} l_i \text{ where } l_i = \{x \in R^2 : L_i(x) = v_i \cdot x = 0\} \right\}$$

and

$$\mathbf{g}_\kappa = \left\{ A \in \mathcal{F}_\kappa^2 : A = \bigcup_{i=1}^{\kappa} l_i \text{ where } l_i = \{x \in R^2 : L_i(x) = v_i \cdot x - \tau_i = 0\} \right\}.$$

Note that $\mathbf{f}_\kappa \cup \mathbf{g}_\kappa = \mathcal{F}_\kappa^d$ and $\mathbf{f}_\kappa \subset \mathbf{g}_\kappa$ for all κ. Therefore, by Theorem 22, $\mathbf{f}_\kappa \preceq_{\mathbf{oc}} \mathbf{g}_\kappa$ for all κ.

Hence, for given κ, the set of ANNs that have a nonzero threshold can solve a classification problem with a higher GC value than the set of ANNs with zero valued thresholds. This is an intuitive demonstration of the use of the ordering $\preceq_{OC}$. Consider Examples 4-7 presented in Section 7.1.2. Note that as the GC value increases from example to example, to implement the dichotomy requires an increasing number of hyperplanes. In other words, κ must increase. Figures 14 and 15 demonstrate the required increase in hyperplanes as the geometry changes. The conclusion would be that if the GC value of a signed set is large, it may be more appropriate to use an ANN with threshold values at each node in the hidden-layer even though thresholds are an additional parameter which must be learned.

7.4 A Comparison of the Ox-Cart Dimension and the V-C Dimension

As mentioned previously, it is not appropriate to compare the *values* of the Ox-Cart Dimension and the V-C Dimension. However, the *mappings* can be compared in the context of the general framework described in Chapter VI. Both mappings are functions of invariants. Both quantifiers are defined based on invariants. V-C dimension is defined on the cardinality invariant which is defined on unsigned sets. The Ox-Cart dimension is defined on the invariant geometric complexity which is defined on signed sets.

Since V-C dimension is defined on arbitrary arrangements of unsigned sets, V-C dimension can only characterize the $\mathcal{F}_\kappa^d$'s ability to solve the worst case geometric arrangement. However, the Ox-Cart dimension is defined more specifically, characterizing $\mathcal{F}_\kappa^d$'s ability to solve any arrangement of signed sets with a given geometric complexity. This is because the invariant, geometric complexity, is more specific than cardinality.

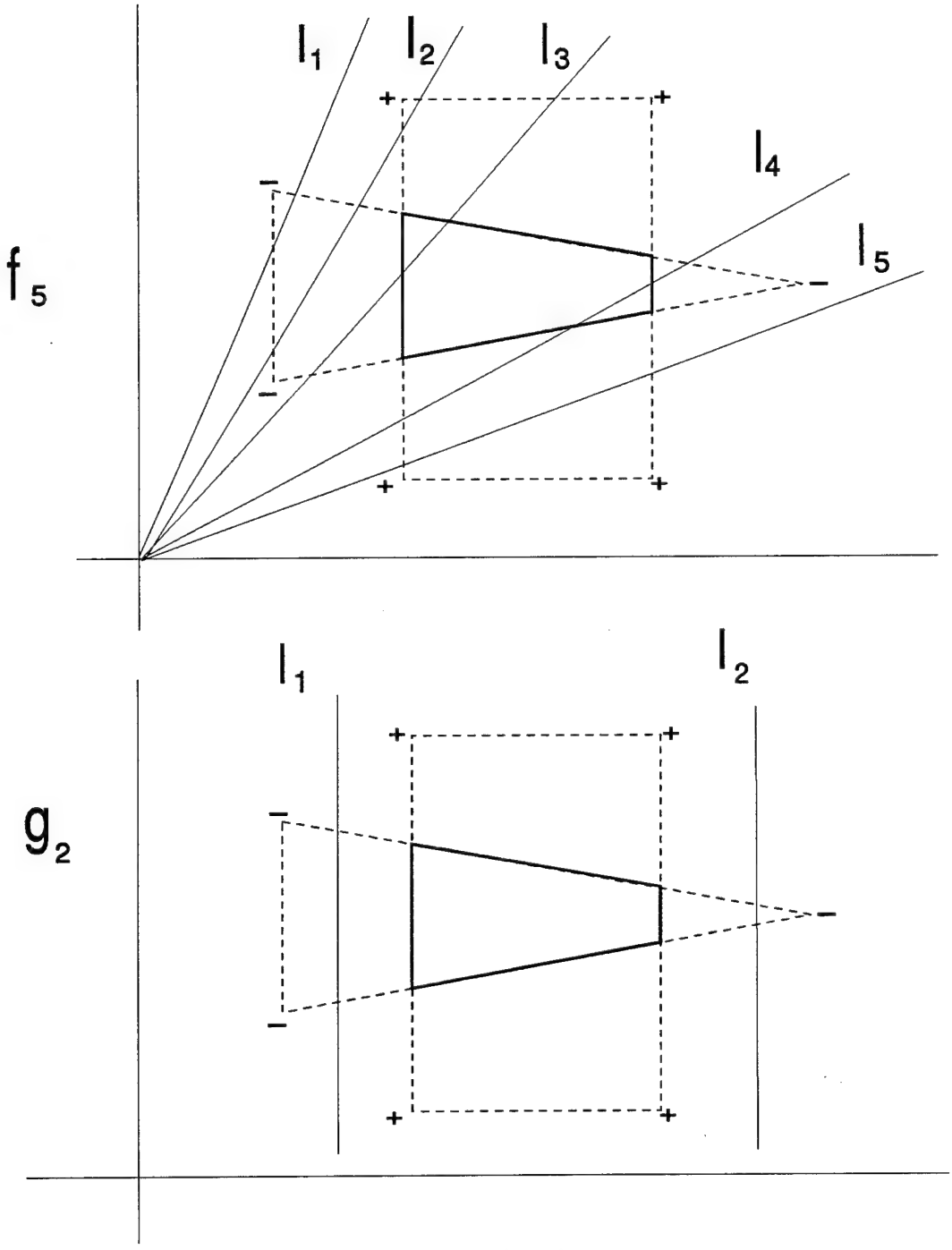


Figure 14. Example 4, solved using $\mathbf{f}_\kappa$ and $\mathbf{g}_\kappa$.

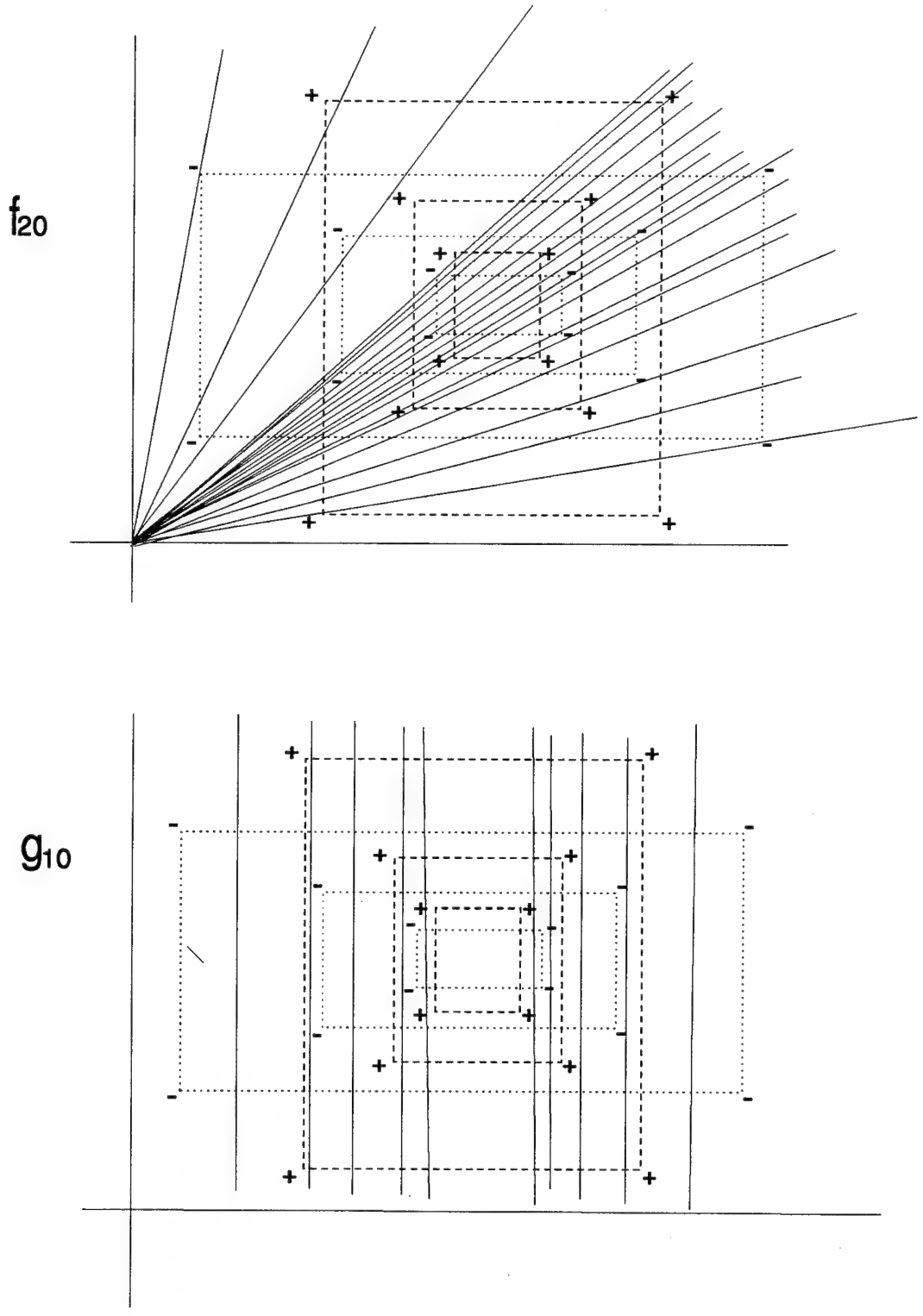


Figure 15. Example 7, solved using $\mathbf{f}_\kappa$ and $\mathbf{g}_\kappa$.

7.5 Conclusions

In summary, this chapter has defined an alternative view of capabilities analysis on ANNs using the generalized framework of Chapter VI. The Ox-Cart dimension was defined based on an invariant, geometric complexity. It was shown that the Ox-Cart dimension induces an ordering that results in a lattice. This lattice is defined as the lattice of feed-forward, single hidden-layer, perceptron, artificial neural networks.

VIII. Follow-On Research

An obvious extension of this research is the investigation of other ANN architectures. Combinatorial Geometry does address arrangements of structures other than hyperplanes. In fact, Grünbaum has provided mathematical background for projecting arbitrary curves into a projective space where they behave as hyperplanes (21). In addition, Zaslavsky defines the geometric lattice $L(A)$ for arrangements of hyperplanes in the projective space with the same ordering as for Euclidean space. Hence, we have the same structure and results available to entities which can be projected into the Euclidean space. Thus this makes the research in this dissertation applicable in a more general setting.

Moreover, once the transformation is made into a projective space, a trade-off analysis can be performed. Applying the quantifiers defined by ν to direct learning algorithms could be a great way to leverage this work. In particular, the quantifiers could be used to compare the power of adding in more complex boundaries, i.e. non-linear decision regions. In other words, the trade-off of additional ANN parameters versus ANN capabilities could be investigated. This extension could lead to more efficient architectures and learning algorithms by continuously evaluating the benefit or need of increasing the complexity of the decision boundaries.

Also, since this research defined the cut-intersection semi-lattice for ANN, much can be discovered about the behavior and structure of the beast. This research addressed specifically the area of determining capabilities. Estimates of computational efficiency can also be derived in the framework of lattice theory.

Currently, the ANN weight space (the set of all vectors that can instantiate a trained ANN) is theoretically assumed to be continuous. In reality, the weight space should be dependent on the resolution of the data. By decreasing the search space to a given resolution, the efficiency of the ANN should increase and this is

something that should be reflected in the mapping that determines the capability. Baum alludes to this fact in his 1987 work (9).

IX. Summary

The intention of this research was to provide a broad base of mathematics on which to study, measure, and improve the capabilities of artificial neural networks. This was accomplished for a specific set of architectures: feed-forward, single hidden-layer, perceptron ANNs. The approach taken for characterizing capabilities was very different from any method described in current literature. The resultant method exhibits new properties that the lack hereof has proved to be the downfall of other methods. Along the way, interesting extensions of old methods were formalized and important links were made between old methods and the new approach proposed in this research.

The approach taken to accomplish the task was to:

1. Understand and clarify current methods;
2. Determine what mathematical structure an ANN exhibits, for example, geometrically, a cut-intersection, semi-lattice was established;
3. Propose new methods and demonstrate their usefulness.

The premise was that measures based on V-C dimension concepts lacked properties that allowed for a meaningful assessment of capabilities. The notion of *shattering* is a hard, strict requirement that actually only measures an ANN's ability to accomplish the absolute worst case of a geometric arrangement of data points. Moreover, the requirement of there being only *at least one* arrangement which can be shattered further limits the ability to generalize capabilities over all possible geometric situations. The reason for this limited approach was that there has been little mathematical structure afforded to ANN architectures formally. When in fact there is a rich set of mathematics (specifically, lattice theory and invariant analysis derived from combinatorial geometry) whose results could be applied directly, once an ANN architecture structure is in place.

The study of current attempts to measure capabilities (Cover, Baum and Sontag) resulted in important extensions to their work. By clarifying the underlying mathematics of their approaches, closed form formulas were derived for evaluating their measures, where before there were only estimates. However, when these concepts were attempted to be generalized, the whole premise of the approach was found to be too limiting. Therefore, the search for an alternative view through combinatorial geometric ideas for measuring capabilities resulted. The following is a compilation of the contributions made by this research.

1. In order to develop the alternative approach of measuring ANN capabilities, the lattice structure of ANN's was established. It was found that a cut-intersection semi-lattice could be defined for ANN's. This result is formalized in Theorem 12.
2. The set of chambers produced by ANNs was shown to be a semi-lattice. This is detailed in Theorem 15.
3. Studying the combinatorial geometry of ANNs produced a formula for the V-C dimension. It was established that the V-C dimension can be reformulated as the well-known geometric invariant of a hyperplane arrangement, cardinality of the set of chambers. This is detailed in Theorem 14.
4. Using the concepts of invariant analysis and lattice theory, a generalized framework in which capability analysis can be performed was described. This resulted in a generalized lattice structure established by Theorem 17. The generalized structure, based on invariant analysis includes a quantifier, partial ordering, and resulting lattice.
5. A new quantifier of ANN capability is defined, called the Ox-Cart dimension. Theorem 21 shows that the Ox-Cart dimension can be used to define the lattice of ANNs.

6. Moreover, the Ox-Cart dimension is based on geometric invariance which is shown to be well-suited for ANN capability analysis since it is directed specifically at signed sets which represent arbitrary classification problems.

9.1 Conclusions

In summary, this document presents an alternative perspective for analyzing the capability of ANNs. The important points are that capability analysis should be viewed through a generalized framework describing the invariant nature of specific problems. That is, the nature of signed sets should be exploited to yield a characterization of separability which can readily be translated to characterize requirements of an ANN architecture designed to separate the signed set. To facilitate this perspective, mathematical structure about ANNs is established. This structure is used to define an invariant called the geometric complexity about signed sets. This is used to define the new ANN capability quantifier called the Ox-Cart dimension.

Bibliography

1. Abbott, James C. *Sets, Lattices and Boolean Algebras.* Allyn and Bacon, Inc., Boston, 1969.
2. Abu-Mostafa, Y. S. "Hints and the VC Dimension," *Neural Computation*, 5: 278-288, (1993).
3. Afarwal, Pankaj K. *Intersection and Decomposition Algorithms for Planar Arrangements*, Cambridge University Press, New York, 1991.
4. Amirikian, Bagrat and Hajime Nishimura, "What Size Network is Good for Generalization of a Specific Task of Interest?", *Neural Networks*, 7: 321-329, (1994).
5. Anderson, Ian. *Combinatorics of Finite Sets.* Oxford Science Publications, Oxford, 1987.
6. Asano, T., J. Hershberger, J. Pach, E. Sontag, D. Souvaine and S. Suri. "Separating Bi-Chromatic Points by Parallel Lines," *Proceedings of the Second Canadian Conference on Computational Geometry*, Ottawa, Canada: 46-49, (1990).
7. Barnsley, Michael. *Fractals Everywhere*, Academic Press, Inc., New York, 1988.
8. Bartlett, P. L. "Vapnik-Chervonenkis Dimension Bounds for Two- and Three-layer Networks," *Neural Computation*, 5: 371-373, (1993).
9. Baum, E. B. "On the Capabilities of Multilayer Perceptrons," *Journal of Complexity*, 4: 193-215, (1988).
10. Baum, E. B., and D. Haussler. "What Size Net Gives Valid Generalization?", *Neural Computation*, 1:151-160, (1989).
11. Benson, Russell, V. *Euclidean Geometry and Convexity*, McGraw-Hill Book Company, New York, 1966.
12. Birkhoff, Garrett. *Lattice Theory*, American Mathematical Society, Providence, 1964.
13. Blumer, A., A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. "Classifying Learnable Geometric Concepts with the Vapnik-Chervonenkis Dimension," *ACM* 0-89791-193-8/86/0500/0273: 273-282, (1986).
14. Blumer, A., A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. "Learnability and the Vapnik-Chervonenkis Dimension," *ACM* 0004-5411: 929-964, (1989).
15. David Cohn and Gerald Tesauro. "How Tight are the Vapnik-Chervonenkis Bounds?" *Neural Computation*, 4: 249-269, (March 1992).
16. Cover, T. M. "Geometry and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," *IEEE Transactions on Electronic Computers*: 326-334, (June 1965).

17. Crapo, Henry H., Gian-Carlo Rota. *On the Foundations of Combinatorial Theory: Combinatorial Geometries*, MIT Press, Mass., 1970.

18. Cybenko, G. "Approximation by Superposition of a Sigmoidal Function," *Math. Control Signals System*, 2: 303-314, (1989).

19. Gallant, A. R., and H. White. "On Learning the Derivatives of an Unknown Mapping With Multilayer Feedforward Networks," *Neural Networks*, 5: 129-138, (1992).

20. Goodman, Jacob E., Richard Pollack, Rephel Wenger, and Tudor Zamfirescu. "Arrangements and Topological Planes", *The American Mathematical Monthly*, Vol. 101, No. 9: 866-878 (November 1994).

21. Grünbaum, B. "Arrangements of Hyperplanes." In: *Proc. Second Louisiana Conf. on Combinatorics and Graph Theory.* Baton Rouge: 44-106 (1971).

22. Halmos, Paul, R. *Measure Theory*, D. Van Norstrand Company, Inc, New York, 1950.

23. Hecht-Nielsen, R. *Neurocomputing*, Addison-Wesley, Menlo Park California, 1990.

24. Hecht-Nielsen, R. "Kolmogorov's Mapping Neural Network Existence Theorem," *Proceedings of the International Conference on Neural Networks*, IEEE: 11-14, (1987), New York.

25. Hoffman, Kenneth and Ray Kunze. *Linear Algebra, Second Edition*, Prentice-Hall, Inc., New Jersey, 1971.

26. Holden, Sean B. *On the Theory of Generalization and Self-Structuring in Linearly Weighted Connectionist Networks.*, Ph.D. thesis, Cambridge University Engineering Department, September 1993. Cambridge Engineering Department Report Number CUED/F-INFENG/TR.161.

27. Hornik, K., M. Stinchcombe, and H. White. "Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks." *Neural Networks*, 3: 551-560, (1990).

28. Hush, D. R., and B. G. Horne. "Progress in Supervised Neural Networks: What's New Since Lippman," *IEEE Signal Processing Magazine*: 8-39, (January 1993).

29. Kung, Joseph P. S. *A Source Book in Matroid Theory*, Birkhäuser, Boston, 1986.

30. Laskowski, M. C. "Vapnik-Chervonenkis Classes of Definable Sets," *J. London Mathematics Society*, 2 (45):377-384, (1992).

31. Luenberger, David G. *Optimization by Vector Space Methods.* John Wiley and Sons Inc., New York, 1969.

32. Lippman, R. P. "An Introduction to Computing with Neural Nets," *IEEE Acoustics, Speech and Signal Processing Magazine*, 4 (2): 4-22, (April 1987).

33. Mehrotra, Kishan G., Chilukuri K. Mohan, and Sanjay Ranka. "Bounds on the Number of Samples Needed for Neural Learning," *IEEE Transactions on Neural Networks*, 2, (6): 548-558 (November 1991).

34. Minsky, M. P., and S. Papert, *Perceptrons*, MIT Press, Cambridge, MA, 1969.

35. Nilsson, N. J. *The Mathematical Foundations of Learning Machines*, Morgan Kaufman Publishers, San Mateo, California, 1990.

36. Orlik Peter, Hiroaki Terao, *Arrangements of Hyperplanes*, Springer-Verlag, New York, 1991.

37. Roberts, S. "On the figures formed by the intercepts of a system of straight lines in a plane, and on analogous relations in space of three dimensions," *Proc. London Math Soc.*, 19: 405-422, (1889).

38. Shonkwiler, R. "Separating the Vertices of N-Cubes by Hyperplanes and its Application to Artificial Neural Networks," *IEEE Transactions on Neural Networks*, 4 (2): 343-347, (March 1993).

39. Sontag, E. D. "Feedforward Nets for Interpolation and Classification," *Journal of Computer and Systems Sciences*, 45: 20-48, (1992).

40. Sontag, E. D. "Sigmoids Distinguish more Efficiently than Heavisides," *Neural Computation*, 1: 470-472, (1989).

41. Takada, Yoshihiro, Xinhua Zhuang, and Hisashi J. Wakita. "A Geometric Algorithm Finding Set of Linear Decision Boundaries," *IEEE Transactions on Signal Processing*, 42, (7): 1887-1891 (July 1994).

42. Tou, J. T., and R. C. Gonzalez. *Pattern Recognition Principles*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1974.

43. Toussaint, Godfried T. "Pattern Recognition and Geometrical Complexity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 5: 1324-1337 (July 1980).

44. Valentine, Frederick A. *Convex Sets*, McGraw-Hill Book Company, New York, 1964.

45. Valiant, L. G. "A Theory of the Learnable," *Communications of the ACM*, 27 (11): 1134-1142, (1984).

46. Vapnik, V. N., and A. Ya. Chervonenkis, "On the Convergence of Relative Frequencies of Events to their Probabilities," *Theory of Probability and its Applications*, 16 (2): 264-280, (1971).

47. Wenocur, R. S. and R. M. Dudley. "Some Special Vapnik-Chervonenkis Classes," *Discrete Mathematics*, 33: 313-318, (1981).

48. Widrow, B. *DARPA Neural Network Study*, AFCEA International Press, Fairfax, VA., 1988.

49. Winder, R. O. "Single Stage Threshold Logic," *Proceedings of the Second Annual Symposium and Papers from the First Annual Symposium on Switching Circuit Theory and Logical Design*, Detroit, Oct. 17-20, 1961 and Chicago, October 9-14, 1960.

50. Zaslavsky, T. "Facing up to Arrangements: Face-Count Formulas for Partitions of Space by Hyperplanes". *Memoirs Amer. Math. Soc.*, 154: 1-95, (1975).

Vita

Martha Ayers Alvey Carter was born in Louisville, Kentucky in 1964. She received a Bachelor of Science in Mathematics in 1986 and a Master of Science in Applied Mathematics in 1987, both from the University of Louisville. She has been employed by the National Air Intelligence Center of the United States Air Force as a mathematician since 1988. Martha married Douglas W. Carter in 1988 and has one child, Gabrielle Suzanne, born in 1994.

Permanent address: 624 Riverwood Dr.
Beavercreek, OH 45430

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 1995	Doctoral Dissertation

4. TITLE AND SUBTITLE
THE MATHEMATICS OF MEASURING ARTIFICIAL NEURAL NETWORKS CAPABILITIES

5. FUNDING NUMBERS

6. AUTHOR(S)
Martha Ayers Alvey Carter, Mathematician, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology, WPAFB OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/DS/ENC/95J-01

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Air Intelligence Center
4115 Hebble Creek Rd, Suite 18, Wright-Patterson AFB OH 45433

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Researchers rely on a measure known as the Vapnik-Chervonenkis (V-C) dimension to capture quantitative measures of capabilities of specific artificial neural network (ANN) architectures. The fundamental thesis of this research is that the V-C dimension is not an appropriate measure of ANN capability. Consequently, the results provide a basis of mathematics on which to build more intuitive measures. Specifically, a cut-intersection semi-lattice is established upon which invariant analysis of an arrangement of hyperplanes can be examined. In addition, a generalized function of invariants is presented as a mechanism for defining capability measures. Moreover, an invariant based on geometric complexity defined by concepts of combinatorial geometry is presented and evaluated. Research on V-C dimension is refined and extended yielding formulas for evaluating V-C dimension for certain cases. As a consequence of the study of combinatorial geometry of hyperplane arrangements, it is shown that the Poincare polynomial also provides an evaluation of V-C dimension of certain ANNs.

14. SUBJECT TERMS	15. NUMBER OF PAGES
artificial neural networks, Vapnik-Chervonenkis dimension, invariant analysis, combinatorial geometry	121
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL